

# Statistical Methods for Machine Learning

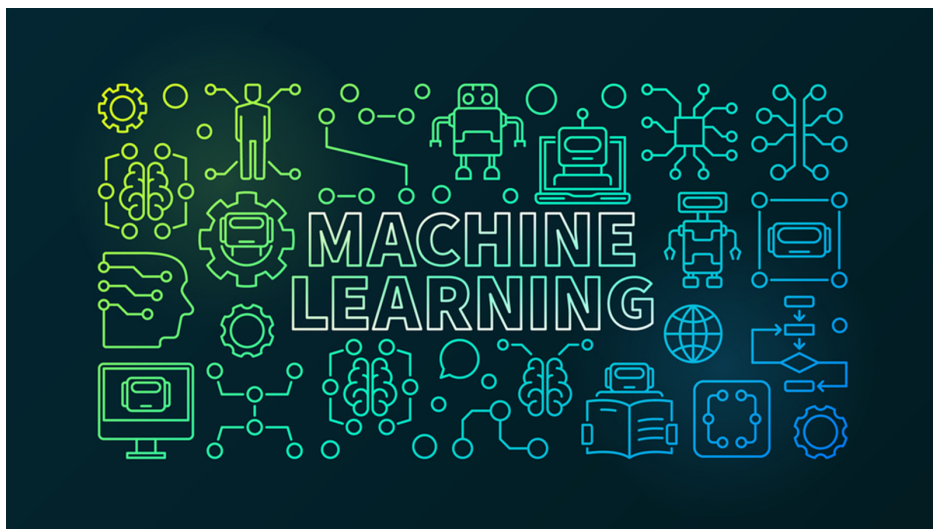


UNIVERSITÀ  
DEGLI STUDI  
DI MILANO

Data Science and Economics  
Università degli Studi di Milano

**Andrea Ierardi**

Lecture notes



# Contents

<b>1</b>	<b>Lecture 1 - 09-03-2020</b>	<b>7</b>
1.1	Introduction of the course . . . . .	7
1.2	Examples . . . . .	7
1.2.1	Spam filtering . . . . .	10
<b>2</b>	<b>Lecture 2 - 10-03-2020</b>	<b>11</b>
2.1	Argomento . . . . .	11
2.2	Loss . . . . .	11
2.2.1	Absolute Loss . . . . .	11
2.2.2	Square Loss . . . . .	12
2.2.3	Example of information of square loss . . . . .	13
2.2.4	labels and losses . . . . .	14
2.2.5	Example TF(idf) documents encoding . . . . .	16
<b>3</b>	<b>Lecture 3 - 16-03-2020</b>	<b>18</b>
3.1	Overfitting . . . . .	20
3.1.1	Noise in the data . . . . .	20
3.2	Underfitting . . . . .	22
3.3	Nearest neighbour . . . . .	22
<b>4</b>	<b>Lecture 4 - 17-03-2020</b>	<b>25</b>
4.1	Computing $h_{NN}$ . . . . .	25
4.2	Tree Predictor . . . . .	27
<b>5</b>	<b>Lecture 5 - 23-03-2020</b>	<b>31</b>
5.1	Tree Classifier . . . . .	31
5.2	Jensen's inequality . . . . .	33
5.3	Tree Predictor . . . . .	37
5.4	Statistical model for Machine Learning . . . . .	38
<b>6</b>	<b>Lecture 6 - 24-03-2020</b>	<b>40</b>
6.1	Bayes Optimal Predictor . . . . .	40
6.1.1	Square Loss . . . . .	41
6.1.2	Zero-one loss for binary classification . . . . .	42
6.2	Bayes Risk . . . . .	45
<b>7</b>	<b>Lecture 7 - 30-03-2020</b>	<b>47</b>
7.1	Chernoff-Hoffding bound . . . . .	47
7.2	Union Bound . . . . .	48
7.3	Studying overfitting of a ERM . . . . .	52

<b>8</b>	<b>Lecture 8 - 31-03-2020</b>	<b>54</b>
8.1	The problem of estimating risk in practise . . . . .	55
8.2	Cross-validation . . . . .	57
8.3	Nested cross validation . . . . .	59
<b>9</b>	<b>Lecture 9 - 06-04-2020</b>	<b>60</b>
9.1	Tree predictors . . . . .	60
9.1.1	Catalan Number . . . . .	62
<b>10</b>	<b>Lecture 10 - 07-04-2020</b>	<b>66</b>
10.1	TO BE DEFINE . . . . .	66
10.2	MANCANO 20 MINUTI DI LEZIONE . . . . .	66
10.3	Compare risk for zero-one loss . . . . .	68
<b>11</b>	<b>Lecture 11 - 20-04-2020</b>	<b>70</b>
11.1	Analysis of $K_{NN}$ . . . . .	70
11.1.1	Study of $K_{NN}$ . . . . .	73
11.1.2	study of trees . . . . .	74
11.2	Non-parametric Algorithms . . . . .	75
11.2.1	Example of parametric algorithms . . . . .	76
<b>12</b>	<b>Lecture 12 - 21-04-2020</b>	<b>77</b>
12.1	Non parametrics algorithms . . . . .	77
12.1.1	Theorem: No free lunch . . . . .	77
12.2	Highly Parametric Learning Algorithm . . . . .	79
12.2.1	Linear Predictors . . . . .	79
12.2.2	MinDisagreement . . . . .	83
<b>13</b>	<b>Lecture 13 - 27-04-2020</b>	<b>84</b>
13.1	Linear prediction . . . . .	84
13.1.1	MinDisOpt . . . . .	84
13.2	The Perception Algorithm . . . . .	87
13.2.1	Perception convergence Theorem . . . . .	88
<b>14</b>	<b>Lecture 14 - 28-04-2020</b>	<b>91</b>
14.1	Linear Regression . . . . .	91
14.1.1	The problem of linear regression . . . . .	91
14.1.2	Ridge regression . . . . .	92
14.2	Percetron . . . . .	93
14.2.1	Online Learning . . . . .	94
14.2.2	Online Gradient Descent (OGD) . . . . .	96

<b>15 Lecture 15 - 04-05-2020</b>	<b>97</b>
15.1 Regret analysis of OGD . . . . .	97
15.1.1 Projected OGD . . . . .	98
<b>16 Lecture 16 - 05-05-2020</b>	<b>102</b>
16.1 Analysis of Perceptron in the non-separable case using OGD framework. . . . .	102
16.1.1 Strongly convex loss functions . . . . .	106
<b>17 Lecture 17 - 11-05-2020</b>	<b>108</b>
17.1 Strongly convex loss functions . . . . .	108
17.1.1 OGD for Strongly Convex losses . . . . .	108
17.1.2 Relate sequential risk and statistical risk . . . . .	109
<b>18 Lecture 18 - 12-05-2020</b>	<b>112</b>
18.1 Kernel functions . . . . .	112
18.1.1 Feature expansion . . . . .	112
18.1.2 Kernels implements feature expansion (Efficiently . . .	113
18.2 Gaussian Kernel . . . . .	114
<b>19 Lecture 19 - 18-05-2020</b>	<b>117</b>
19.1 Support Vector Machine (SVM) . . . . .	120
<b>20 Lecture 20 - 19-05-2020</b>	<b>121</b>

# List of Figures

2.1	Example of domain of $K_{NN}$	12
2.2	Example of domain of $K_{NN}$	12
3.1	Example of domain of $K_{NN}$	18
3.2	Example of domain of $K_{NN}$	23
3.3	Example of domain of $K_{NN}$	23
4.1	Example of domain of $K_{NN}$	26
4.2	Example of domain of $K_{NN}$	27
4.3	Example of domain of $K_{NN}$	28
4.4	Example of domain of $K_{NN}$	29
4.5	Example of domain of $K_{NN}$	29
4.6	Example of domain of $K_{NN}$	29
4.7	Example of domain of $K_{NN}$	30
5.1	Example of domain of $K_{NN}$	31
5.2	Example of domain of $K_{NN}$	32
5.3	Example of domain of $K_{NN}$	33
5.4	Example of domain of $K_{NN}$	33
5.5	Example of domain of $K_{NN}$	34
5.6	Example of domain of $K_{NN}$	35
5.7	Example of domain of $K_{NN}$	36
5.8	Example of domain of $K_{NN}$	36
5.9	Example of domain of $K_{NN}$	36
5.10	Example of domain of $K_{NN}$	37
6.1	Example of domain of $K_{NN}$	41
6.2	Example of domain of $K_{NN}$	43
6.3	Example of domain of $K_{NN}$	43
6.4	Example of domain of $K_{NN}$	44
6.5	Example of domain of $K_{NN}$	45
7.1	Example	48
7.2	Example	49
7.3	Example	49
7.4	Example	50
7.5	Example	50
7.6	Draw of how $\hat{h}$ , $h^*$ and $f^*$ are represented	51
8.1	Representation of $\hat{h}$ , $h^*$ and $f^*$	54
8.2	Example	55

8.3	Splitting test and training set . . . . .	57
8.4	K-folds . . . . .	58
8.5	Nested Cross Validation . . . . .	59
9.1	Tree building . . . . .	60
9.2	Tree with at most N node . . . . .	61
9.3	Algorithm for tree predictors . . . . .	64
10.1	Point (2) - where $y = cx + q$ $y = -cx + q$ . . . . .	68
10.2	Point . . . . .	69
11.1	Example of domain of $K_{NN}$ . . . . .	70
11.2	Diagonal length . . . . .	71
11.3	Shape of the function . . . . .	72
11.4	Parametric and non parametric growing as training set getting larger . . . . .	76
12.1	Tree building . . . . .	78
12.2	Dot product . . . . .	79
12.3	Dot product . . . . .	80
12.4	Hyperplane . . . . .	80
12.5	Hyperplane . . . . .	81
12.6	Hyperplane . . . . .	81
12.7	Example of one dimensional hyperplane . . . . .	82
13.1	Tree building . . . . .	85
13.2	Tree building . . . . .	85
13.3	Tree building . . . . .	86
13.4	Feasibility problem . . . . .	86
13.5	. . . . .	87
13.6	. . . . .	88
14.1	. . . . .	92
14.2	. . . . .	94
14.3	. . . . .	95
14.4	. . . . .	96
15.1	. . . . .	98
15.2	. . . . .	98
15.3	. . . . .	99
15.4	. . . . .	100
16.1	. . . . .	102

16.2 Hinge loss . . . . .	103
16.3 . . . . .	104
16.4 Example of more type of convex function . . . . .	106
18.1 . . . . .	112
18.2 . . . . .	115
18.3 . . . . .	115
19.1 Draw of SVG . . . . .	120

# Lecture 1 - 09-03-2020

## 1.1 Introduction of the course

In this course we look at the principle behind design of Machine learning. Not just coding but have an idea of algorithm that can work with the data.

We have to fix a mathematical framework: some statistic and mathematics.

### Work on ML on a higher level

ML is data inference: make prediction about the future using data about the past

- Clustering → grouping according to similarity
- Planning → (robot to learn to interact in a certain environment)
- Classification → (assign meaning to data) example: Spam filtering  
I want to predict the outcome of this individual or i want to predict whether a person click or not in a certain advertisement.

## 1.2 Examples

Classify data into categories:

- Medical diagnosis: data are medical records and categories are diseases
- Document analysis: data are texts and categories are topics
- Image analysts: data are digital images and for categories name of objects in the image (but could be different).
- Spam filtering: data are emails, categories are spam vs non spam.
- Advertising prediction: data are features of web site visitors and categories could be click/non click on banners.

Classification : **Different from clustering** since we do not have semantically classification (spam or not spam) → like meaning of the image.  
I have a semantic label.



Clustering: i want to group data with similarity function.

Planning: Learning what to do next

Clustering: Learn similarity function

Classification: Learn semantic labels meaning of data

Planning: Learn actions given state

In classification is an easier than planning task since I'm able to make prediction telling what is the semantic label that goes with data points.

If i can do classification i can clustering.

If you do planning you probably classify (since you understanding meaning in your position) and then you can also do clustering probably.

We will focus on classification because many tasks are about classification.

Classify data in categories we can image a set of categories.

For instance the tasks:

'predict income of a person'

'Predict tomorrow price for a stock'

The label is a number and not an abstract thing.

We can distinguish two cases:

- The label set  $\rightarrow$  set of possible categories for each data point. For each of this could be finite set of abstract symbols (case of document classification, medical diagnosis). So the task is classification.
- Real number (no bound on how many of them). My prediction will be a real number and is not a category. In this case we talk about a task of regression.

Classification: task we want to give a label predefined point in abstract categories (like YES or NO)

Regression: task we want to give label to data points but this label are numbers.

When we say prediction task: used both for classification and regression tasks.

Supervised learning: Label attached to data (classification, regression)

Unsupervised learning: No labels attached to data (clustering)

In unsupervised the mathematical modelling and way algorithm are score and can learn from mistakes is a little bit harder. Problem of clustering is harder to model mathematically.

You can cast planning as supervised learning: i can show the robot which is the right action to do in that state. But that depends on planning task is formalised.

Planning is higher level of learning since include task of supervised and unsupervised learning.

Why is this important ?

Algorithm has to know how to given the label.

In ML we want to teach the algorithm to perform prediction correctly. Initially algorithm will make mistakes in classifying data. We want to tell algorithm that classification was wrong and just want to perform a score. Like giving a grade to the algorithm to understand if it did bad or really bad. So we have mistakes!

Algorithm predicts and something makes a mistake  $\rightarrow$  we can correct it.

Then algorithm can be more precisely. We have to define this mistake.

Mistakes in case of classification:

- If category is the wrong one (in the simple case). We have a binary signal where we know that category is wrong.

How to communicate it?

We can use the loss function: we can tell the algorithm whether is wrong or not.

**Loss function:** measure discrepancy between 'true' label and predicted label.

So we may assume that every datapoint has a true label. If we have a set of topic this is the true topic that document is talking about. It is typical in supervised learning.

How good the algorithm did?

$$\ell(y, \hat{y}) \leq 0$$

were  $y$  is true label and  $\hat{y}$  is predicted label

We want to build a spam filter where 0 is not spam and 1 is spam and that's a Classification task:

$$\ell(y, \hat{y}) = \begin{cases} 0, & \text{if } \hat{y} = y \\ 1, & \text{if } \hat{y} \neq y \end{cases}$$

**The loss function is the “interface” between algorithm and data.**

So algorithm know about the data through the loss function.

If we give a useless loss function the algorithm will not perform good: is important to have a good loss function.

### 1.2.1 Spam filtering

$Y = \{spam, no\ spam\}$

Binary classification  $|Y| = 2$

We have two main mistake:

- False positive:  $y = \text{non spam}, \hat{y} = \text{spam}$
- False negative:  $y = \text{spam}, \hat{y} = \text{no spam}$

It is the same mistake? No if i have important email and you classify as spam that's bad and if you show me a spam than it's ok.

So we have to assign a different weight.

$$\ell(y, \hat{y}) = \begin{cases} 2 & \text{if } FP \\ 1 & \text{if } FN \\ 0 & \text{otherwise} \end{cases}$$

**We have to take more attention on positive mistake**

Even in binary classification, mistakes are not equal.

# Lecture 2 - 10-03-2020

## 2.1 Argomento

Classification tasks

Semantic label space  $Y$

Categorization  $Y$  finite and

small Regression  $Y$  appartiene ad  $\mathbb{R}$

How to predict labels?

Using the lost function  $\rightarrow ..$

Binary classification

Label space is  $Y = -1, +1$

Zero-one loss

$$\ell(y, \hat{y}) = \begin{cases} 0, & \text{if } \hat{y} = y \\ 1, & \text{if } \hat{y} \neq y \end{cases}$$

*FP*  $\hat{y} = 1, \quad y = -1$

*FN*  $\hat{y} = -1, \quad y = 1$

Losses for regression?

$y$ , and  $\hat{y} \in \mathbb{R}$ ,

so they are numbers!

One example of loss is the absolute loss: absolute difference between numbers

## 2.2 Loss

### 2.2.1 Absolute Loss

$$\ell(y, \hat{y}) = |y - \hat{y}| \Rightarrow \text{absolute loss}$$

Some inconvenient properties:

- ...
- Derivative only two values (not much informations)

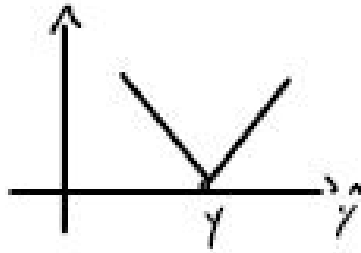


Figure 2.1: Example of domain of  $K_{NN}$

### 2.2.2 Square Loss

$$\ell(y, \hat{y}) = (y - \hat{y})^2 \Rightarrow \text{square loss}$$

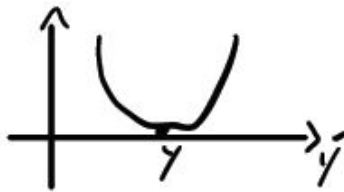


Figure 2.2: Example of domain of  $K_{NN}$

Derivative :

- more informative
- and differentiable

Real numbers as label  $\rightarrow$  regression.

Whenever taking difference between two prediction make sense (value are numbers) then we are talking about regression problem.

Classification as categorization when we have small finite set.

### 2.2.3 Example of information of square loss

$$\begin{aligned}\ell(y, \hat{y}) &= (y - \hat{y})^2 = F(y) \\ F'(\hat{y}) &= -2 \cdot (y - \hat{y})\end{aligned}$$

- I'm under sho or over and how much
- How much far away from the truth

$$\ell(y, \hat{y}) = |y - \hat{y}| = F(y') \cdot F'(y) = \text{Sign}(y - \hat{y})$$

Question about the future

Will it rain tomorrow?

We have a label and this is a binary classification problem.

My label space will be  $Y = \text{"rain"}, \text{"no rain"}$

We don't get a binary prediction, we need another space called prediction space (or decision space).

$$Z = [0, 1]$$

$\hat{y} \in Z$       $\hat{y}$  is my prediction of rain tomorrow

$\hat{y} = \mathbb{P}(y = \text{"rain"})$       $\rightarrow$  my guess is tomorrow will rain (not sure)

$$y \in Y \quad \hat{y} \in Z$$

*quad*How can we manage loss?

*Put numbers in our space*

$\{1, 0\}$      where 1 is rain and 0 no rain

I measure how much I'm far from reality.

So loss behave like this and the punishment is gonna go linearly??

26..

However is pretty annoying. Sometime I prefer to punish more so i going quadratically instead of linearly.

There are other way to punish this.

I called **logarithmic loss**

We are extending a lot the range of our loss function.

$$\ell(y, \hat{y}) = |y - \hat{y}| \in [0, 1] \quad \ell(y, \hat{y}) = (y - \hat{y})^2 \in [0, 1]$$

If i want to expand the punishment i use logarithmic loss

$$\ell(y, \hat{y}) = \begin{cases} \ln \frac{1}{\hat{y}}, & \text{if } y = 1(\text{rain}) \\ \ln \frac{1}{1-\hat{y}}, & \text{if } y = 0(\text{no rain}) \end{cases}$$

$F(\hat{y}) \rightarrow$  can be 0 if i predict with certainty

If  $\hat{y} = 0.5$   $\ell(y, \frac{1}{2}) = \ln 2$  constant losses in each prediction

$$\lim_{\hat{y} \rightarrow 0^+} \ell(1, \hat{y}) = +\infty$$

We give a vanishing probability not rain but tomorrow will rain.

So this is  $+\infty$

$$\lim_{\hat{y} \rightarrow 1^-} \ell(0, \hat{y}) = +\infty$$

The algorithm will be punish high more the prediction is not real. Algorithm will not get 0 and 1 because for example is impossible to get a perfect prediction.

This loss is useful to give this information to the algorithm.

Now we talk about labels and losses

## 2.2.4 labels and losses

Data points: they have some semantic labels that denote some true about this data points and we want to predict this labels.

We need to define what data points are: number? Strings? File? Typically they are stored in database records

They can have very precise structure or more homogeneously structured

A data point can be viewed as a vector in some d dimensional real space. So it's a vector of number

$$\mathbb{R}^d X = (x_1, x_2, \dots, x_d) \in \mathbb{R}^c$$

Image can be viewed as a vector of pixel values (grey scale 0-255).

I can use geometry to learn because point are in my Euclidean space. Data can be represented as point in Euclidean space. Images are list of pixel that are pretty much the same range and structure (from 0 to 255). It's very natural to put them in a space.

Assume X can be a record with heterogeneous fields:

For example medical records, we have several values and each fields has his meaning by it's own. (Sex, weight, height, age, zip code)

Each one has a different range, in some cases is numerical but something have like age ..

Does have any sense to see a medical record as a point since coordinates have different meaning.

**Fields are not comparable.**

This is something that you do: when you want to solve some inference you have to decide which are the label and what is the label space and we have to encode the data points.

Data algorithm expect some homogenous interface. In this case algorithm has to build records with different values of fields.

This is something that we have to pay attention too.

You can always each range of values in number. So ages is number, sex you can give 0 and 1, weight number and zip code is number.

How ever geometry doesn't make sense since I cannot compare this coordinates.

Linear space i can sum up as vector: i can make linear combination of vectors.

Inner product to measure angles! (We will see in linear classifier).

I can scramble the number of my zip code.

So we get problems with sex and zip code

Why do we care about geometry? I can use geometry to learn.

However there is more to that, geometry will carry some semantically information that I'm going to preserve during prediction.

I want to encode my images as vectors in a space. Images with dog.....

PCA doesn't work because assume we encode in linear space.

We hope geometry will help us to predict label correctly and sometimes i hard to convert data into geometry point.

Example of comparable data: images, or documents.

Assume we have documents with corpus (set of documents).

Maybe in English and talk about different thing and different words.

X is a document and i want to encode X into a point fix in bidimensional space.

There is a way to encode a set of documents in point in a fixed dimensional space in such way it make sense this coordinate are comparable.

I can represent fields with  $[0,1]$  for Neural network for example. But they have no geometrical meaning



## 2.2.5 Example TF(idf) documents encoding

TF encoding of docs.

1. Extract where all the words from docs
2. Normalize words (nouns, adjectives, verbs ...)
3. Build a dictionary of normalized words

Doc  $x = (x_1, \dots, x_d)$

I associate a coordinate for each word in a dictionary.

$d$  = number of words in dictionary

I can decide that

$x_i = 1$      *If  $i$ -th word of dictionary occurs in doc.*

$x_i = 0$      *Else*

$X_i$    *number of time  $i$ -th word occur in doc.*

Longer documents will have higher value of coordinates that are not zero.

Now i can do the TF encoding in which  $x_i$  = frequency with which  $i$ -th word occur in dictionary.

You cannot sum dog and cat but we are considering them frequencies so we are summing frequency of words.

This encoding works well in real words.

I can choose different way of encoding my data and sometime i can encode a real vector

I want

1. A predictor  $f : X \rightarrow Y$  (in weather  $X \rightarrow Z$ )
2.  $X$  is our data space (where points live)
3.  $X = \mathbb{R}^d$  images
4.  $X = X_1 x \dots x X_d$  Medical record
5.  $\hat{y} = f(x)$  predictor for  $X$

$(x, y)$

We want to predict a label that is much closer to our label. How?

Loss function: so this is my setting and is called an example.

Data point together with label is a “example”

We can get collection of example making measurements or asking people. So we can always recover the true label.

We want to replace this process with a predictor (so we don't have to bored a person).

$y$  is the ground truth for  $x \rightarrow$  mean reality!

If i want to predict stock for tomorrow, i will wait tomorrow to see the ground truth.

# Lecture 3 - 16-03-2020

Data point  $x$  represented as sequences of measurement and we called this measurements features or attributes.

$$x = (x_1, \dots, x_d) \quad x_1 \text{ feature value } x \in X^d \quad X = \mathbb{R}^d \quad X = X_1 \cdot x \dots X_d \cdot x$$

Label space  $Y$

Predictor  $f : X \rightarrow Y$

Example  $(x, y)$   $y$  is the label associated with  $x$   
 ( $\rightarrow y$  is the correct label, the ground truth)

Learning with example  $(x_1, y_1) \dots (x_m, y_m)$  training set

Training set is a set of examples with every algorithm can learn.....

Learning algorithm take training set as input and produces a predictor as output.



Figure 3.1: Example of domain of  $K_{NN}$

With image recognition we use as measurement pixels.

How do we measure the power of a predictor?

A learning algorithm will look at training set, algorithm and generate the predictor. Now the problem is verify the score.

Now we can consider a test set collection of example

$$\text{Test set} \quad (x'_1, y'_1) \dots (x'_n, y'_n)$$

Typically we collect big dataset and then we split in training set and test set randomly.

### Training and test are typically disjoint

How we measure the score of a predictor? We compute the average loss.

The error is the average loss in the element in the test set.

$$\text{Test error} \quad \frac{1}{n} \cdot \sum_{t=1}^n \ell(f(x'_t), y'_t)$$

In order to simulate we collect the test set and take the average loss of the predictor of the test set. This will give us idea of how the..

Proportion of test and train depends in how big the dataset is in general.

Our **Goal**: A learning algorithm 'A' must output f with a small test error.

A does not have access to the test set. (Test set is not part of input of A).

Now we can think in general on how a learning algorithm should be design.

We have a training set so algorithm can say:

**'A' may choose f based on performance on training set.**

$$\text{Training error} \quad \hat{\ell}(f) = \frac{1}{m} \cdot \sum_{t=1}^m \ell(f(x_t), y_t)$$

Given the training set  $(x_1, \dots, x_m)(y_1, \dots, y_m)$

If  $\hat{\ell}(f)$  for same f, then test of f is also small

Fix F set of predictors output  $\hat{f}$

$$\hat{f} = \arg \min_{f \in F} \hat{\ell}(f)$$

### This algorithm is called Empirical Risk Minimiser (ERM)

When this strategy (ERM) fails?

ERM may fails if for the given training set there are:

Many  $f \in F$  with small  $\hat{\ell}(f)$ , but not all of them have small test error

There could be many predictor with small error but some of them may have big test error. Predictor with the smallest training error doesn't mean we will have the smallest test error.

I would like to pick  $f^*$  such that:

$$f^* = \arg \min_{f \in F} \frac{1}{n} \cdot \sum_{t=1}^n \ell(f(x'_t), y'_t)$$

where  $\ell(f(x'_t), y'_t)$  is the test error

ERM works if  $f^*$  such that  $f^* = \arg \min_{f \in F} \hat{\ell}(f)$

So minimising training and test????? Check videolecture

We can think of  $f$  as finite since we are working on a finite computer.  
We want to see why this can happen and we want to formalise a model in which we can avoid this to happen by design: We want when we run ERM choosing a good predictor with ..... PD

## 3.1 Overfitting

We called this as overfitting: specific situation in which 'A' (where A is the learning algorithm) overfits if  $f$  output by A tends to have a training error much smaller than the test error.

A is not doing his job (outputting large test error) this happen because test error is misleading.

Minimising training error doesn't mean minimising test error. Overfitting is bad.

Why this happens?

This happen because we have **noise in the data**

### 3.1.1 Noise in the data

Noise in the data:  $y_t$  is not deterministically associated with  $x_i$ .

Could be that datapoint appears more times in the same test set. Same datapoint is repeated actually I'm mislead since training and dataset not coincide. Minimising the training error can take me away from the point that minimise the test error.

Why this is the case?

- Some **human in the loop**: label assigned by people.(Like image contains certain object but human are not objective and people may have different opinion)
- **Lack of information**: in weather prediction i want to predict weather error. Weather is determined by a large complicated system. If i have humidity today is difficult to say for sure that tomorrow will rain.

When data are not noise i should be ok.

**Labels are not noisy**

Fix test set and training set.

$$\begin{aligned} \exists f^* \in F \quad y'_t = f^*(x'_t) \quad \forall (x'_t, y'_t) \quad \text{in test set} \\ y_t = f^+(x_t) \quad \forall (x_t, y_t) \quad \text{in training set} \end{aligned}$$

Think a problem in which we have 5 data points(vectors) :

$\vec{x}_1, \dots, \vec{x}_5$  in some space  $X$

We have a binary classification problem  $Y = \{0, 1\}$

$\{\vec{x}_1, \dots, \vec{x}_5\} \in X \quad Y = \{0, 1\}$

$F$  contains all possible classifiers  $2^5 = 32 \quad f : \{x_1, \dots, x_5\} \rightarrow \{0, 1\}$

Example					
	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$f$	0	0	0	0	0
$f'$	0	0	0	0	1
$f''$	..	..	..	..	..

Training set  $x_1, x_2, x_3 \quad f^+$

Test set  $x_4, x_5 \quad f^*$

4 classifier  $f \in F$  will have  $\hat{\ell}(f) = 0$

$(x_1, 0) \quad (x_2, 1) \quad (x_3, 0)$

$(x_4, ?) \quad (x_5, ?)$

$f^*(x_4) \quad f^*(x_5)$

If not noise i will have deterministic data but in this example (worst case) we get problem.

I have 32 classifier to choose: i need a larger training set since i can't distinguish predictor with small and larger training(?) error. So overfitting noisy or can happen with no noisy but few point in the dataset to define which predictor is good.

## 3.2 Underfitting

'A' underfits when  $f$  output by A has training error close to test error but they are both large.

Close error test and training error is good but they are both large.

$A \equiv \text{ERM}$ , then A underfits if  $F$  is too small  $\rightarrow$  not containing too much predictors

In general, given a certain training set size:

- Overfitting when  $|F|$  is too large (not enough points in training set)
- Underfitting when  $|F|$  is too small

Proportion predictors and training set

$|F|$ , i need  $\ln|F|$  bits of info to uniquely determine  $f^* \in F$   
 $m \gg \ln|F|$  when  $|F| < \infty$  where  $m$  is the size of training set

## 3.3 Nearest neighbour

This is completely different from ERM and is one of the first learning algorithms. This exploits the geometry of the data. Assume that our data space  $X$  is:

$$X \equiv \mathbb{R}^d \quad x = (x_1, \dots, x_d) \quad y \in \{-1, 1\}$$

$S$  is the training set  $(x_1, y_1) \dots (x_m, y_m)$

$$x_t \in \mathbb{R}^d \quad y_t \in \{-1, 1\}$$

$d = 2 \rightarrow 2$ -dimensional vector

where + and - are labels

### Point of test set

If i want to predict this point?

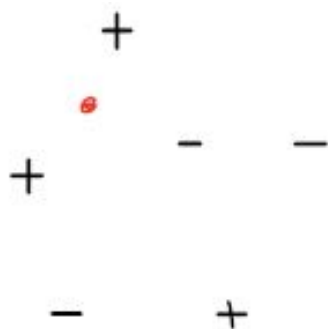


Figure 3.2: Example of domain of  $K_{NN}$

Maybe if point is close to point with label i know then. Maybe they have the same label.

$\hat{y} = +$  or  $\hat{y} = -$

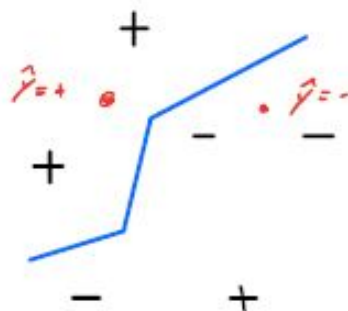


Figure 3.3: Example of domain of  $K_{NN}$

I can came up with some sort of classifier.

Given  $S$  training set, i can define  $h_{NN} X \rightarrow \{-1, 1\}$

$h_{NN}(x) = \text{label } y_t \text{ of the point } x_t \text{ in } S \text{ closest to } X$

**(the breaking rule for ties)**

For the closest we mean euclidian distance



$$X = \mathbb{R}^d$$

$$\|x - x_t\| = \sqrt{\sum_{e=1}^d (x_e - x_t, e)^2}$$

$$\hat{\ell}(h_{NN}) = 0$$

$$h_{NN}(x_t) = y_t$$

**training error is 0!**

# Lecture 4 - 17-03-2020

We spoke about Knn classifier with voronoi diagram

$$\hat{\ell}(h_{NN}) = 0 \quad \forall \text{ Training set}$$

$h_{NN}$  predictor needs to store entire dataset.

## 4.1 Computing $h_{NN}$

Computing  $h_{NN}(x)$  requires computing distances between  $x$  and points in the training set.

$$\Theta(d) \quad \text{time for each distance}$$

NN  $\rightarrow$  1-NN

We can generalise NN in K-NN with  $k = 1, 3, 5, 7$  so odd  $K$

$h_{k-NN}(x)$  = label corresponding to the majority of labels of the  $k$  closest point to  $x$  in the training set.

How big could  $K$  be if i have  $n$  point?

I look at the  $k$  closest point

When  $k = m$ ?

The majority, will be a constant classifier  $h_{k-NN}$  is constant and corresponds to the majority of training labels

Training error is always 0 for  $h_{NN}$ , while for  $h_{k-NN}$  will be typically  $> 0$ , with  $k > 1$

Image: one dimensional classifier and training set is repeated. Is the plot of 1-NN classifier.

Positive and negative.  $K = 1$  error is 0.

In the second line we switch to  $k = 3$ . Second point doesn't switch and third will be classify to positive and we have training mistake.

Switches corresponds to border of voronoi partition.

$K_{NN}$  For multiclass classification

( $|Y| > 2$ ) for regression  $Y \equiv \mathbb{R}$

Average of labels of  $K$  neighbours  $\rightarrow$  i will get a number with prediction.

I can weight average by distance

You can vary this algorithm as you want.

Let's go back to Binary classification.

The  $k$  parameter is the effect of making the structure of classifier more complex and less complex for small value of  $k$ .

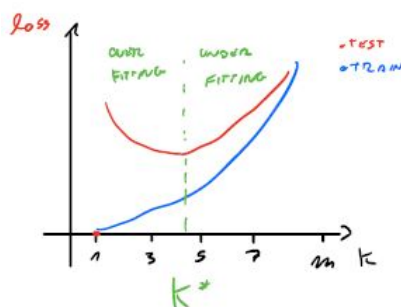


Figure 4.1: Example of domain of  $K_{NN}$

Fix training set and test set

Accuracy as oppose to the error

Show a plot. Training error is 0 at  $k = 0$ .

As i go further training error is higher and test error goes down. At some point after which training and set met and then after that training and test error goes up (accuracy goes down).

If i run algorithm is going to be overfitting: training error and test error is high and also underfitting since testing and training are close and both high. Trade off point is the point in  $x = 23$  (more or less).

There are some heuristic to run NN algorithm without value of  $k$ .

## History

- $K_{NN}$ : from 1960  $\rightarrow X \equiv \mathbb{R}^d$
- Tree predictor: from 1980

## 4.2 Tree Predictor

If a give you data not welled defined in a Euclidean space.

$X = X_1 \cdot x \cdot \dots \cdot X_d \cdot x$  Medical Record

$X_1 = \{Male, Female\}$

$X_2 = \{Yes, No\}$

so we have different data

I want to avoid comparing  $x_i$  with  $x_j$ ,  $i \neq j$

so comparing different feature and we want to compare each feature with each self. I don't want to mix them up.

We can use a tree!

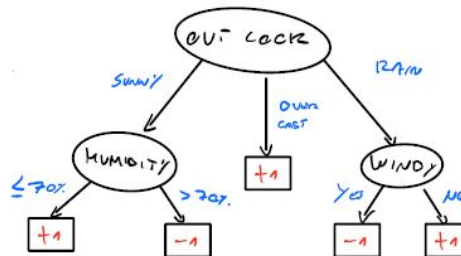


Figure 4.2: Example of domain of  $K_{NN}$

I have 3 features:

- outlook =  $\{sunny, overcast, rain\}$
- humidity =  $\{[0, 100]\}$
- windy =  $\{yes, no\}$

Tree is a natural way of doing decision and abstraction of decision process of one person. It is a good way to deal with categorical variables.

What kind of tree we are talking about?

Tree has inner node and leaves. Leaves are associated with labels ( $Y$ ) and inner nodes are associated with test.

- Inner node  $\rightarrow$  test
- Leaves  $\rightarrow$  label in  $Y$

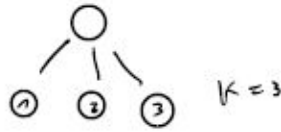


Figure 4.3: Example of domain of  $K_{NN}$

Test if a function  $f$  (NOT A PREDICTOR!)

Test  $f_i X_i \rightarrow \{1, \dots, k\}$

where  $k$  is the number of children (inner node) to which test is assigned

In a tree predictor we have:

- Root node
- Children are ordered (i know the order of each branch that come out from the node)

$X = \{Sunny, 50\%, No\} \rightarrow$  are the parameters for  $\{outlook, humidity, windy\}$

$$f_i = \begin{cases} 1, & \text{if } x_2 \in [30\%, 60\%] \\ 2, & \text{if } otherwise \end{cases}$$

where the numbers 1 and 2 are the children

A test is partitioning the range of values of a certain attribute in a number of elements equal to number of children of of the node to which the test is assigned.

$h_T(x)$  is always the label of a leaf of T

This leaf is the leaf to which  $x$  is **routed**

Data space for this problem (outlook,..) is partitioned in the leaves of the tree. It won't be like voronoi graph. How do I build a tree given a training set? How do i learn a tree predictor given a training set?

- Decide tree structure (how • many node, leaves ecc..)
- Decide test on inner nodes
- Decide labels on leaves

We have to do this all together and process will be more dynamic. For simplicity binary classification and fix two children for each inner node.

$Y = \{-1, +1\}$

2 children for each inner node

What's the simplest way?  
 Initial tree and correspond to a constant classifier



Figure 4.4: Example of domain of  $K_{NN}$

Majority of all example



Figure 4.5: Example of domain of  $K_{NN}$

$(x_1, y_1) \dots (x_m, y_m)$   
 $x_t \in X \quad y_t \in \{-1, +1\}$   
 Training set  $S = \{(x, y) \in S, x \text{ is routed to } \ell\}$   
 $S_\ell^+$

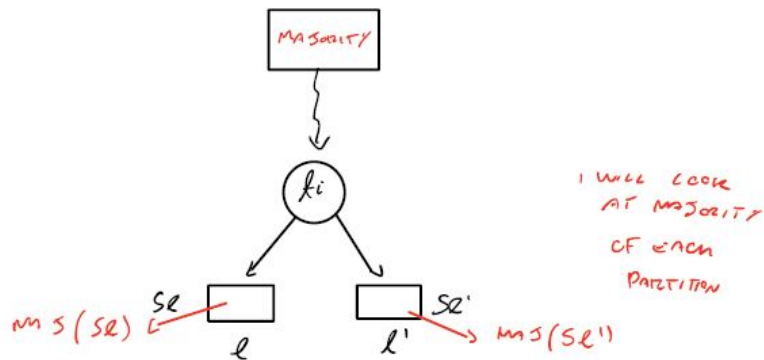


Figure 4.6: Example of domain of  $K_{NN}$

$S_\ell$  and  $S_{\ell'}$  are given by the result of the test, not the labels and  $\ell$  and  $\ell'$ .

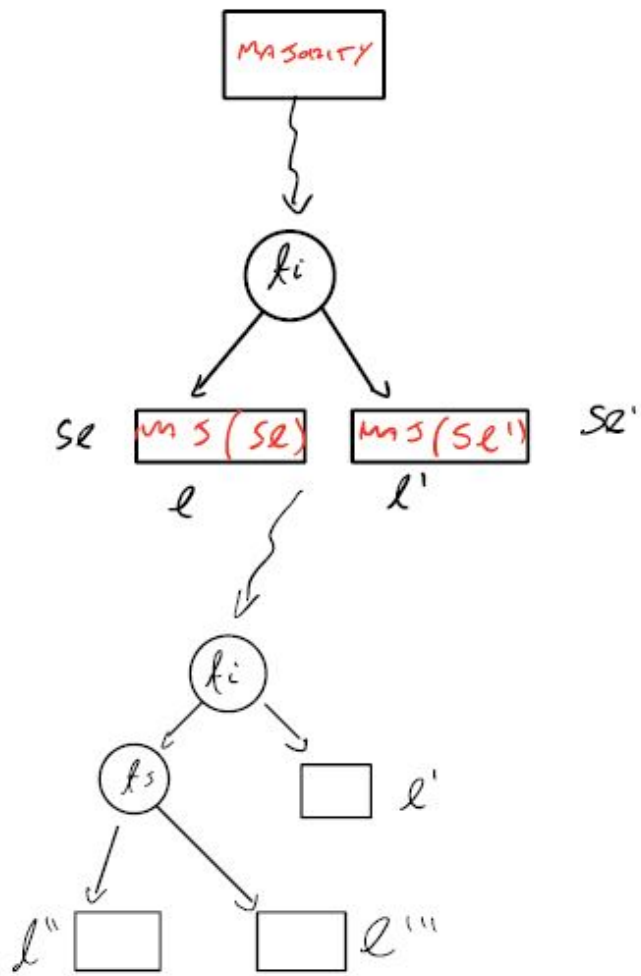


Figure 4.7: Example of domain of  $K_{NN}$

# Lecture 5 - 23-03-2020

## 5.1 Tree Classifier

Supposed we groped a tree up to this point and we are wandering how to grow it.

$S$  Training set  $(x_1, y_1) \dots (x_m, y_m)$ ,  $x_1 \in X$

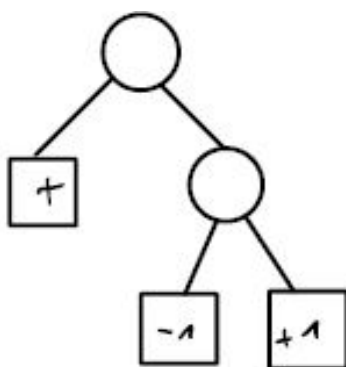


Figure 5.1: Example of domain of  $K_{NN}$

$$S_\ell \equiv \{(x_1, y_1) \mid x_t \text{ is router to } \ell\}$$

$$y_1 \in \{-1, 1\}$$

$$S_\ell^+ \equiv \{(x_1, y_1) \in S_\ell : y_t = +1\}$$

$$S_\ell^- \equiv \{(x_1, y_1) \in S_\ell : y_t = -1\} \quad S_\ell^+ \cap S_\ell^- \equiv \emptyset \quad S_\ell \equiv S_\ell^+ \cup S_\ell^-$$

$$N_\ell = |S_\ell| \quad N_\ell^+ = |S_\ell^+| \quad N_\ell^- = |S_\ell^-|$$

$$N_\ell = N_\ell^- + N_\ell^+$$

leaf  $\ell$  classifies all training example  $(S_\ell)$

$$Y_\ell = \begin{cases} +1, & \text{If } N_\ell^+ \geq N_\ell^- \\ -1, & \text{If otherwise} \end{cases}$$



$\ell$  makes a mistake on  $\min\{N_\ell^+, N_\ell^-\}$  example in  $S_\ell$

$$\begin{aligned} \hat{\ell}(h_T) &= \frac{1}{m} \cdot \sum_{\ell} \min\left\{\frac{N_\ell^+}{N_\ell}, \frac{N_\ell^-}{N_\ell}\right\} \cdot N_\ell = \\ &= \frac{1}{m} \cdot \sum_{\ell} \psi \cdot \left(\frac{N_\ell^+}{N_\ell}\right) \cdot N_\ell \quad \longrightarrow \quad \frac{N_\ell^+}{N_\ell} = 1 - \frac{N_\ell^-}{N_\ell} \end{aligned}$$

where  $\psi(a) = \min\{a, 1 - a\}$   $a \in [0, 1]$

I want to replace inner node with other leaves.

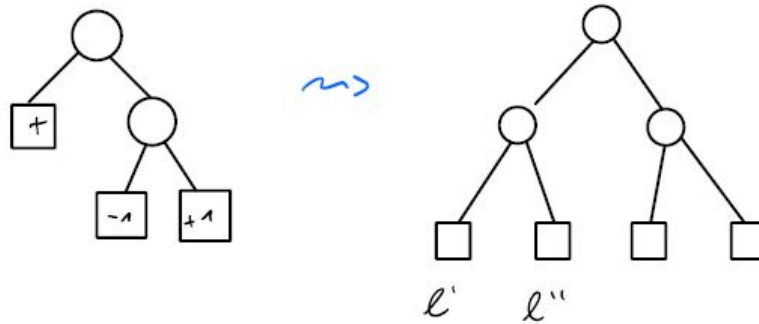


Figure 5.2: Example of domain of  $K_{NN}$

How is training error going to change? (when i replace inner nodes with other leaves)

I'm hoping my algorithm is not going to overfit (if training error goes to 0 also testing error goes to 0).

## 5.2 Jensen's inequality

If  $\psi$  is a concave function  $\rightarrow$  (like  $\log$  or  $\sqrt{\cdot}$  )

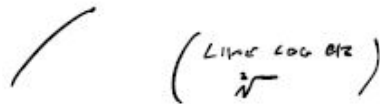


Figure 5.3: Example of domain of  $K_{NN}$

Also  $\psi$  is a function that map 0 to 1,  $\rightarrow \psi [0, 1] \rightarrow \mathbb{R}$

$\psi(\alpha \cdot a + (1-\alpha) \cdot b) \geq \alpha \cdot \psi(a) + (1-\alpha) \cdot \psi(b)$  Also 2° derivative is negative

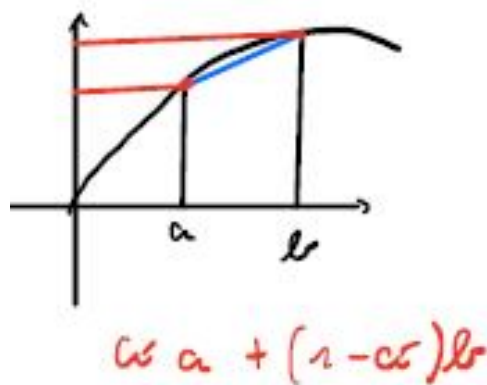


Figure 5.4: Example of domain of  $K_{NN}$

$$\hat{\ell}(h_T) = \frac{1}{m} \cdot \sum_{\ell} \psi\left(\frac{N_{\ell}^+}{N_{\ell}}\right) \cdot N_{\ell}$$

Look a single contribution fo a leaf  $\ell$  to training error

$$\psi\left(\frac{N_{\ell}^+}{N_{\ell}}\right) \cdot N_{\ell} = \psi\left(\frac{N_{\ell}'^+}{N_{\ell}'} \cdot \frac{N_{\ell}'}{N_{\ell}} + \frac{N_{\ell}''^+}{N_{\ell}''} \cdot \frac{N_{\ell}''}{N_{\ell}}\right) \cdot N_{\ell}$$

where  $\frac{N_{\ell}'}{N_{\ell}} = \alpha$  and  $\frac{N_{\ell}''}{N_{\ell}} = 1 - \alpha$  so  $\frac{N_{\ell}'}{N_{\ell}} + \frac{N_{\ell}''}{N_{\ell}} = 1 \rightarrow \alpha + 1 - \alpha = 1$

$$N_{\ell'}^+ + N_{\ell''}^+ = N_\ell$$

I want to check function *min* concave between 0 and 1.

$$\min(0, 1) = 0 \quad \psi(a) = \min(\alpha, 1 - \alpha)$$

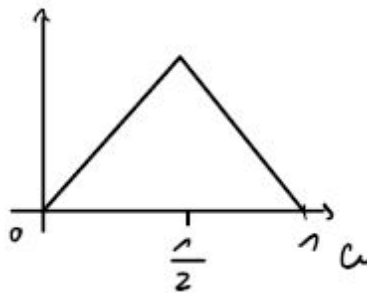


Figure 5.5: Example of domain of  $K_{NN}$

This is a concave function and now I can apply Jensen's inequality

$$\begin{aligned} \psi\left(\frac{N_\ell^+}{N_\ell}\right) \cdot N_\ell &\geq \left(\frac{N_{\ell'}^+}{N_\ell} \cdot \psi\left(\frac{N_{\ell'}^+}{N_{\ell'}^+}\right) + \frac{N_{\ell''}^+}{N_\ell} \cdot \psi\left(\frac{N_{\ell''}^+}{N_{\ell''}^+}\right)\right) \cdot N_\ell = \\ &= \boxed{\psi\left(\frac{N_{\ell'}^+}{N_{\ell'}^+}\right) \cdot N_{\ell'}^+ + \psi\left(\frac{N_{\ell''}^+}{N_{\ell''}^+}\right) \cdot N_{\ell''}^+} \end{aligned}$$

This are the contribution of  $\ell'$  and  $\ell''$  to the training error

Every time i split my tree my training error is never going to increase since we have a concave function.

Whenever I'm growing my tree training error is going to be smaller.

**Every time a leaf is expanded the training error never goes up. (Hopelly will go down)**

I'll should always grow the tree by expanding leave that decrease the training error as much as possible.

If i take the effort of growing the tree i should get benefits. I can imaging that if i grow the tree at random my training error is going to drop down error (but maybe will derive overfitting).

For now is just an intuition since we will introduced statistical learning model.

Could be complicated and tree big may have 100 leaves and there could be many ways of associating a test with those leaves.

I can spend a lot of time to select which leaf is the best promising to split.

- Grow the tree by expanding leaf that decrease the training error as much as possible
- In general we can assume:  
greedy algorithm at each step pick the pair leaf and test that cause (approximative) the largest decrease in training error

In practice we want to optimize this all the way since it's time expensive. That's the approximately since we are not every time sure.

— MANCA PARTE —

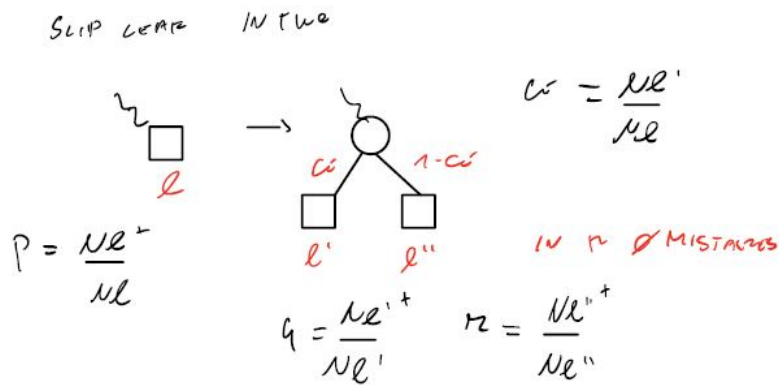


Figure 5.6: Example of domain of  $K_{NN}$

$$p = 0.8 \quad q = 1 \quad r = 1 \quad \alpha = 60\%$$

Net Change in number of mistakes

$$\psi(p) - (\alpha \cdot \psi(q) + (1 - \alpha) \cdot \psi(r)) =$$

$$l - l' + l''$$

Fraction of example miss classified  $l$  - error  $l'$  + error  $l''$

$$= 0.2 - \left(\frac{1}{2} \cdot 0.4 + \frac{1}{2} \cdot 0\right) = 0$$

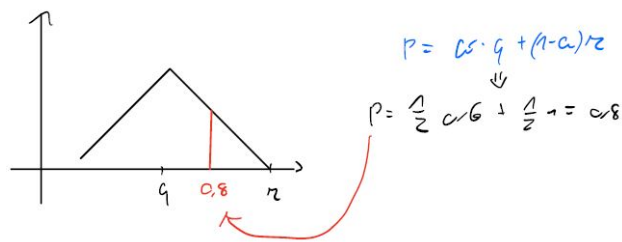


Figure 5.7: Example of domain of  $K_{NN}$

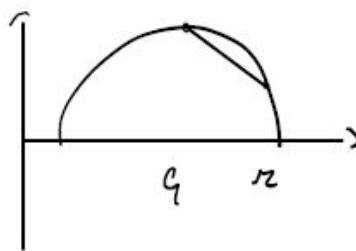


Figure 5.8: Example of domain of  $K_{NN}$

Idea is to replace minimum function with convex combination.

$$\psi(\alpha) = \min \{ \alpha, 1 - \alpha \} \quad \psi(a) \geq \psi(\alpha)$$

$$\begin{cases} \psi_1(\alpha) = 2 \cdot \alpha \cdot (1 - \alpha) \longrightarrow \text{GNI} \\ \psi_2(\alpha) = -\frac{\alpha}{2} \cdot \ln \alpha - \frac{1-\alpha}{2} \cdot \ln(1 - \alpha) \longrightarrow \text{ENTROPY} \\ \psi_3(\alpha) = \sqrt{\alpha \cdot (1 - \alpha)} \end{cases}$$

All this functions has this shape (concave???)



Figure 5.9: Example of domain of  $K_{NN}$

In practise Machine Learning algorithm use GNI or entropy to control the split

## 5.3 Tree Predictor

- Multi class classification  $|Y| > 2 \rightarrow$  take majority
- Regression  $Y = \mathbb{R} \rightarrow$  take average of labels in  $S_\ell$

I still take majority among different classes.

Take average of labels in  $S_\ell$

Unless  $\frac{N_\ell^+}{N_\ell} \in 0, 1 \quad \forall$  leaves  $\ell, \hat{\ell}(h_T) > 0$

Unless leaves are "pured", the training error will be bigger than 0.

In general, i can always write  $\hat{\ell}(h_t)$  to 0 by growing enough the tree unless there are  $x_1$  in the Time Series such that  $(x_t, y_t)(x_t, y'_t)$  with  $y_t \neq y'_t$  both occur.

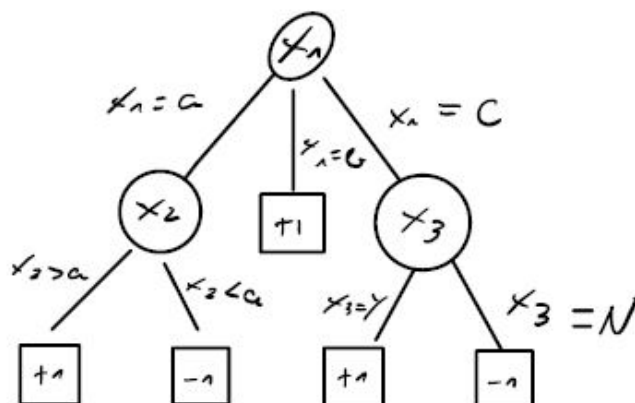


Figure 5.10: Example of domain of  $K_{NN}$

$$if(x_1 = \alpha) \wedge (x_2 \geq \alpha) \vee (x_1 = b) \vee (x_1 = c) \wedge (x_3 = y)$$

then predict 1

else

then predict -1

— Picture of tree classifier of iris dataset. —

I'm using due attribute at the time.

Each data point is a flower and i can measure how petal and sepal are long.

I can use two attribute and i test this two. I can see the plot of the tree classifier (second one) making test splitting data space into region that has this sort of “blackish” shape ( like boxes: blue box, red box, yellow box)  
 A good exercise in which I want to reconstruct the tree given this picture.

## 5.4 Statistical model for Machine Learning

To understand Tree classifier, nearest neighbour and other algorithm...  
 It's important to understand that the only way to have a guideline in which model to choose.

**This mathematical model are developed to learning and choose learning algorithm.**

Now let start with theoretical model.

- How example  $(x, y)$  are generated to create test set and training set?  
 We get the dataset but we need to have a mathematical model for this process.  $(x, y)$  are drawn from a fixed but unknown probability distribution on the pairs  $X$  and  $Y$  ( $X$  data space,  $Y$  label set o label space)
- Why  $X$  should be random?  
 In general we assumed that not all the  $x$  in  $X$  are equally likely to be observed. I have some distribution over my data point and this said that I'm most like to get a datapoint to another.
- How much label?  
 Often labels are not determined uniquely by their datapoints because labels are given by human that have their subjective thoughts and also natural phenomena. Labels are stochastic phenomena given a data-point: i will have a distribution.

We're going to write (in capital)  $(X, Y)$  since they are random variable drawn from  $D$  on  $X \cdot Y$  A dataset  $(X_1, Y_1) \dots (X_m, Y_m)$  they are drawn independently from  $D$  (distribution on examples)

When I get a training the abstraction of process collecting a training set  $D$  is a joint probability distribution over  $X \cdot Y$   
 where  $D_x$  is the marginal over  $X \rightarrow D_y|x$  (conditional of  $Y$  given  $X$ ).

I can divided my draw in two part. I draw sample and label from conditional.??

Any dataset ( training or test ) is a random sample (campione casuale) in the statistical sense  $\rightarrow$  so we can use all stastical tools to make inference.



# Lecture 6 - 24-03-2020

$(X, Y)$  We random variables drawn iid from  $D$  on  $X \cdot Y \longrightarrow$  where  $D$  is fixed but unknown

Independence does not hold. We do not collect datapoints to an independent process.

Example: identify new article and i want to put categories. The feed is highly depend on what is happening in the world and there are some news highly correlated. Why do we make an assumption that follows reality? Is very convenient in mathematical term. If you assume Independence you can make a lot of process in mathematical term in making the algorithm.

If you have enough data they look independent enough. Statistical learning is not the only way of analyse algorithms  $\longrightarrow$  we will see in linear ML algorithm and at the end you can use both statistical model s

## 6.1 Bayes Optimal Predictor

$$f^* : X \rightarrow Y$$

$$f^*(x) = \operatorname{argmin} \mathbb{E}[\ell(y, \hat{y})|X = x] \quad \hat{y} \in Y$$

In general  $Y$  given  $X$  has distribution  $D_y|X = x$

Clearly  $\forall h : X \rightarrow Y$

$$\mathbb{E}[\ell(y, f^*(x))|X = x] \leq \mathbb{E}[\ell(y, h(x))|X = x]$$

$$X, Y \quad \mathbb{E}[Y|X = x] = F(x) \quad \longrightarrow \text{Conditional Expectation}$$

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}(Y)$$

Now take Expectation for distribution

$$\mathbb{E}[\ell(y, f^*(x))] \leq [\mathbb{E}(\ell(y, h(x)))]$$

where **risk is smaller in  $f^*$**

I can look at the quantity before

$l_d$  Bayes risk  $\longrightarrow$  Smallest possible risk given a learning problem

$$l_d(f^*) > 0 \quad \text{because } y \text{ are still stochastic given } X$$

Learning problem can be complem  $\rightarrow$  large risk

### 6.1.1 Square Loss

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$

I want to compute bayes optimal predictor  
 $\hat{y}, y \in \mathbb{R}$

$$f^*(x) = \operatorname{argmin}_{\hat{y} \in \mathbb{R}} \mathbb{E} [(y - \hat{y})^2 | X = x]$$

we use  $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = \operatorname{argmin}_{\hat{y}} \mathbb{E} [y^2 + \hat{y}^2 - 2 \cdot y \cdot \hat{y} | X = x] =$

Dropping  $y^2$  i remove something that is not important for  $\hat{y}$

$$\begin{aligned} &= \operatorname{argmin}(\mathbb{E} [y^2 | X = x] + \hat{y}^2 - 2 \cdot \hat{y} \cdot \mathbb{E} [y | X = x]) = \\ &= \operatorname{argmin}(\hat{y}^2 - 2 \cdot \hat{y} \cdot \mathbb{E} [y | X = x]) = \end{aligned}$$

Expectation is a number, so it's a **constant**

Assume  $\square = y^2$

$$\operatorname{argmin} [\square + \hat{y}^2 + 2 \cdot \hat{y} \cdot \mathbb{E} [Y | X = x]]$$

where  $G(\hat{y})$  is equal to the part between [...]

$$\frac{dG(\hat{y})}{d\hat{y}} = 2 \cdot \hat{y} - 2 \cdot \mathbb{E} [y | X = x] = 0 \quad \longrightarrow \quad \text{So setting derivative to 0}$$

Suppose we have a learning domain



Figure 6.1: Example of domain of  $K_{NN}$

$$G'(\hat{y}) = \hat{y}^2 - 2 \cdot b \cdot \hat{y}$$

$$\hat{y} = \mathbb{E} [y | X = x] \quad f^*(x) = \mathbb{E} [y | X = x]$$

Square loss is nice because expected prediction is ...

In order to predict the best possible we have to estimate the value given data point.

$$\begin{aligned} & \mathbb{E} [(y - f^*(x))^2 | X = x] = \\ & = \mathbb{E} [(y - \mathbb{E}[y|X = x])^2 | X = x] = \text{Var} [Y|X = x] \end{aligned}$$

### 6.1.2 Zero-one loss for binary classification

$$Y = \{-1, 1\}$$

$$\ell(y, \hat{y}) = I\{\hat{y} \neq y\} \quad I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

If  $\hat{y} \neq y$  true, indicator function will give us 1, otherwise it will give 0

$$D \text{ on } X \cdot Y \quad D_x^* \quad D_{y|x} = D$$

$$D_x \quad \eta : X \longrightarrow [0, 1] \quad \eta = \mathbb{P}(y = 1 | X = x)$$

$$D \rightsquigarrow (D_x, \eta) \longrightarrow \text{Distribution 0-1 loss}$$

$X \sim D_x \longrightarrow$  Where  $\sim$  mean "draw from" and  $D_x$  is marginal distribution

$$Y = 1 \quad \text{with probability } \eta(x)$$

$$D_{y|x} = \{\eta(x), 1 - \eta(x)\}$$

Suppose we have a learning domain

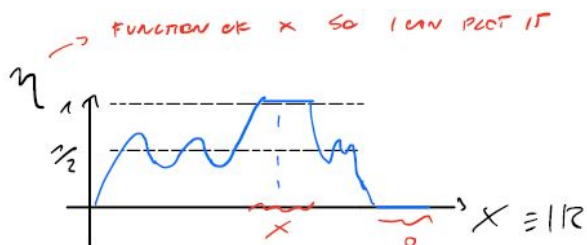


Figure 6.2: Example of domain of  $K_{NN}$

where  $\eta$  is a function of  $x$ , so i can plot it  
 $\eta$  will te me  $Prob(x) =$   
 $\eta$  tells me a lot how hard is learning problem in the domain  
 $\eta(x)$  is not necessary continous



Figure 6.3: Example of domain of  $K_{NN}$

$\eta(x) \in \{0, 1\}$       $y$  is always determined by  $x$

How to get  $f^*$  from the graph?

$$f^+ : X \rightarrow \{-1, 1\}$$

$$Y = \{-1, +1\}$$

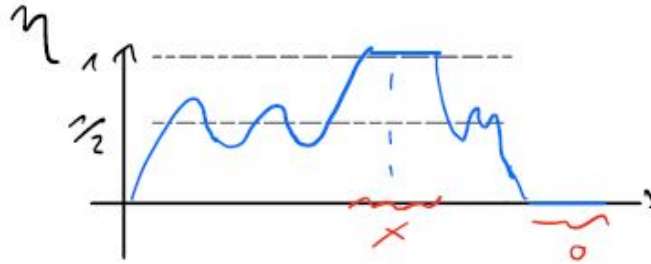


Figure 6.4: Example of domain of  $K_{NN}$

=====  
 MANCA ROBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  
 =====

$$f^*(x) = \operatorname{argmin} \mathbb{E} [\ell(y, \hat{y}) | X = x] = \rightarrow \hat{y} \in \{-1, +1\}$$

$$= \operatorname{argmin} \mathbb{E} [I\{\hat{y} = 1\} \cdot I\{Y = -1\} + I\{\hat{y} = -1\} \cdot I\{y = 1\} | X = x] =$$

we are splitting wrong cases

$$= \operatorname{argmin} ( I\{\hat{y} = 1\} \cdot \mathbb{E} [ I\{Y = -1\} | X = x ] + I\{\hat{y} = -1\} \cdot \mathbb{E} [ I\{y = 1\} | X = x ] ) = \quad *$$

We know that:

$$\mathbb{E} [ I\{y = -1\} | X = x ] = 1 \cdot \mathbb{P}(\hat{y} = -1 | X = x) + 0 \cdot \mathbb{P}(y = 1 | X = x) =$$

$$\mathbb{P}(x = -1 | X = x) = 1 - \eta(x)$$

$$* = \operatorname{argmin} ( I\{\hat{y} = 1\} \cdot (1 - \eta(x)) + I\{\hat{y} = -1\} \cdot (\eta(x)) )$$

where Blue colored  $I\{\dots\} = 1^\circ$  and Orange  $I\{\dots\} = 2^\circ$

I have to choose -1 or +1 so we will **remove one of the two** ( $1^\circ$  or  $2^\circ$ )

It depend on  $\eta(x)$ :

- If  $\eta(x) < \frac{1}{2}$   $\longrightarrow$  kill 1°
- Else  $\eta(x) \geq \frac{1}{2}$   $\longrightarrow$  kill 2°

$$f^*(x) = \begin{cases} +1 & \text{if } \eta(x) \geq \frac{1}{2} \\ -1 & \text{if } \eta(x) < \frac{1}{2} \end{cases}$$

## 6.2 Bayes Risk

$$\mathbb{E}[I\{y \neq f^*(x)\} | X = x] = \mathbb{P}(y \neq f^*(x) | X = x)$$

$$\eta(x) \geq \frac{1}{2} \Rightarrow \hat{y} = 1 \Rightarrow \mathbb{P}(y \neq 1 | X = x) = 1 - \eta(x)$$

$$\eta(x) < \frac{1}{2} \Rightarrow \hat{y} = -1 \Rightarrow \mathbb{P}(y \neq -1 | X = x) = \eta(x)$$

Conditiona risk for 0-1 loss is:

$$\begin{aligned} \mathbb{E}[\ell(y, f^*(x)) | X = x] &= I\{\eta(x) \geq \frac{1}{2}\} \cdot (1 - \eta(x)) + I\{\eta(x) < \frac{1}{2}\} \cdot \eta(x) = \\ &= \min\{\eta(x), 1 - \eta(x)\} \end{aligned}$$

$$\mathbb{E}[\ell, f^*(x)] = \mathbb{E}[\min\{\eta(x), 1 - \eta(x)\}]$$

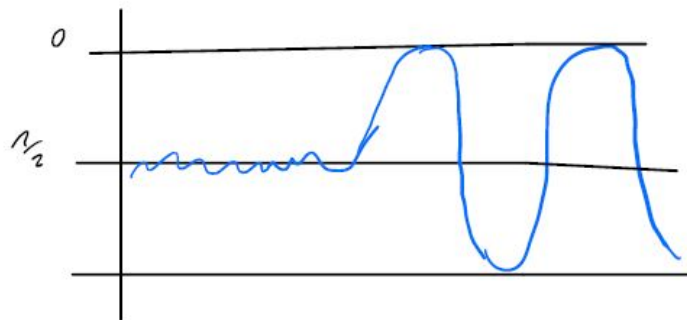


Figure 6.5: Example of domain of  $K_{NN}$

Conditional risk will be high around the half so min between the two is around

the half since the labels are random i will get an error near 50%.  
My condition risk will be 0 in the region in the bottom since label are going to be deterministic.

# Lecture 7 - 30-03-2020

Bounding statistical risk of a predictor

Design a learning algorithm that predict with small statistical risk

$$(D, \ell) \quad \ell_d(h) = \mathbb{E}[\ell(y), h(x)]$$

where  $D$  is unknown

$$\ell(y, \hat{y}) \in [0, 1] \quad \forall y, \hat{y} \in Y$$

We cannot compute statistical risk of all predictor.

We assume statistical loss is bounded so between 0 and 1. Not true for all losses (like logarithmic).

Before design a learning algorithm with lowest risk, How can we estimate risk?

We can use test error  $\rightarrow$  way to measure performances of a predictor  $h$ . We want to link test error and risk.

Test set  $S' = \{(x'_1, y'_1) \dots (x'_n, y'_n)\}$  is a random sample from  $D$

How can we use this assumption?

Go back to the definition of test error

Sample mean (IT: Media campionaria)

$$\hat{\ell}_s(h) = \frac{1}{n} \cdot \sum_{t=1}^n \ell(\hat{y}_t, h(x'_t))$$

i can look at this as a random variable  $\ell(y'_t, h(x'_t))$

$$\mathbb{E}[\ell(y'_t, h(x'_t))] = \ell_D(h) \rightarrow \text{risk}$$

Using law of large number (LLN), i know that:

$$\hat{\ell} \rightarrow \ell_D(h) \quad \text{as } n \rightarrow \infty$$

We cannot have a sample of  $n = \infty$  so we will introduce another assumption: the **Chernoff-Hoffding bound**

## 7.1 Chernoff-Hoffding bound

$$Z_1, \dots, Z_n \quad \text{iid random variable} \quad \mathbb{E}[Z_t] = u$$

all drawn for the same distribution

$$t = 1, \dots, n \quad \text{and} \quad 0 \leq Z_t \leq 1 \quad t = 1, \dots, n \quad \text{then} \quad \forall \varepsilon > 0$$



$$\mathbb{P}\left(\frac{1}{n} \cdot \sum_{t=1}^n z_t > u + \varepsilon\right) \leq e^{-2\varepsilon^2 n} \quad \text{or} \quad \mathbb{P}\left(\frac{1}{n} \cdot \sum_{t=1}^n z_t < u - \varepsilon\right) \leq e^{-2\varepsilon^2 n}$$

as sample size then  $\downarrow$

$$Z_t = \ell(Y'_t, h(X'_t)) \in [0, 1]$$

$(X'_1, Y'_1) \dots (X'_n, Y'_n)$  are *iid* therefore,

$\ell(Y'_t, h(X'_t)) \quad t = 1, \dots, n$  are also *iid*

We are using the bound of  $e$  to bound the deviation of this.

## 7.2 Union Bound

Union bound: a collection of event not necessary disjoint, then i know that probability of the union of this event is the at most the sum of the probabilities of individual events

$$A_1, \dots, A_n \quad \mathbb{P}(A_1 \cup \dots \cup A_n) \leq \sum_{t=1}^n \mathbb{P}(A_t)$$

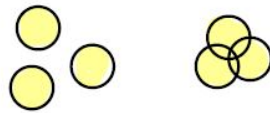


Figure 7.1: Example

that's why  $\leq$

$$\mathbb{P}\left(|\hat{\ell}_{s'}(h) - \ell_D(h)| > \varepsilon\right)$$

This is the probability according to the random draw of the test set.

If test error differ from the risk by a number  $\varepsilon > 0$ . I want to bound the probability. This two thing will differ by more than  $\varepsilon$ . How can i use the Chernoff bound?

$$|\hat{\ell}_{s'}(h) - \ell_D(h)| > \varepsilon \quad \Rightarrow \quad \hat{\ell}_{s'}(h) - \ell_D(h) > \varepsilon \quad \vee \quad \hat{\ell}_D(h) - \ell_{s'}(h) > \varepsilon$$



Figure 7.2: Example

$$A, B \quad A \Rightarrow B \quad \mathbb{P}(A) < \mathbb{P}(B)$$

$$\begin{aligned} \mathbb{P}\left(|\hat{\ell}_{s'}(h) - \ell_D(h)| > \varepsilon\right) &\leq \mathbb{P}\left(|\hat{\ell}_{s'}(h) - \ell_D(h)|\right) \cup \mathbb{P}\left(|\hat{\ell}_D(h) - \ell_{s'}(h)|\right) \leq \\ &\leq \mathbb{P}\left(\hat{\ell}_{s'} > \ell_D(h) + \varepsilon\right) + \mathbb{P}\left(\hat{\ell}_{s'} < \ell_D(h) - \varepsilon\right) \leq 2 \cdot e^{-2\varepsilon^2 n} \Rightarrow \text{we call it } \delta \\ \varepsilon &= \sqrt{\frac{1}{2 \cdot n} \ln \frac{2}{\delta}} \end{aligned}$$

The two events are disjoint

This mean that probability of this deviation is at least delta!

$$|\hat{\ell}_{s'}(h) - \ell_D(h)| \leq \sqrt{\frac{1}{2 \cdot n} \ln \frac{2}{\delta}} \quad \text{with probability at least } 1 - \delta$$

Test error of true estimate is going to be good for this value ( $\delta$ )

Confidence interval for risk at confidence level 1-delta.

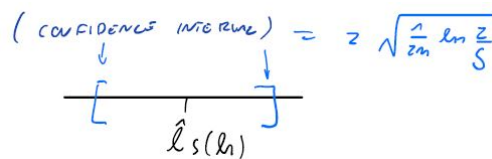


Figure 7.3: Example

I want to take  $\delta = 0,05$  so that  $1 - \delta$  is 95%. So test error is going to be an estimate of the true risk which is precise that depend on how big is the test set ( $n$ ).

As  $n$  grows I can pin down the position of the true risk.

This is how we can use probability to make sense of what we do in practise.

If we take a predictor  $h$  we can compute the risk error estimate.

We can measure how accurate is our risk error estimate.

**Test error is an estimate of risk for a given predictor (h).**

$$\mathbb{E}[\ell(Y'_t, h(X'_t))] = \ell_D(h)$$

**h is fixed with respect to S'**  $\rightarrow$  h does not depend on the test set. So learning algorithm which produce h not have access to test set.

If we use test set we break down this equation.

Now, how to **build a good algorithm?**

Training set  $S = \{(x_1, y_1) \dots (x_m, y_m)\}$  random sample

$A(S) = h$  predictor output by A given S where A is **learning algorithm as function of training set S.**

$$\forall S \quad A(S) \in H \quad h^* \in H$$

$\ell_D(h^*) = \min \ell_D(h)$   $\hat{\ell}_s(h^*)$  is closed to  $\ell_D(h^*) \rightarrow$  **it is going to have small error**

where  $\ell_D(h^*)$  is the **training error of  $h^*$**

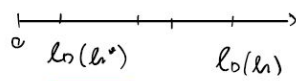


Figure 7.4: Example

This guy  $\ell_D(h^*)$  is closest to 0 since optimum

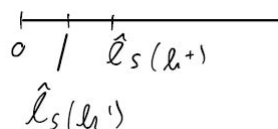


Figure 7.5: Example

In risk we get opt in  $h^*$  but in empirical one we could get another  $h'$  better than  $h^+$

In order to fix on a concrete algorithm we are going to take the empirical Islam minimiser (ERM) algorithm.

A is *ERM* on H  $(A) = \hat{h} = (\epsilon) \operatorname{argmin} \hat{\ell}_S(h)$

Once I pick  $\hat{h}$  i can look at training error of ERM

$$\hat{\ell}_S(\hat{h}) \text{ of } \hat{h} = A(S)$$

where  $\hat{\ell}_S$  is the training error

Should  $\hat{\ell}_S(\hat{h})$  be close to  $\ell_D(\hat{h})$  ?

I'm interested in empirical error minimiser and do a trivial decomposition.

$$\begin{aligned} \ell_d(\hat{h}) = & \ell_D(\hat{h}) - \ell_d(h^*) + && \longrightarrow \text{Variance error} \Rightarrow \text{Overfitting} \\ & + \ell_d(h^*) - \ell_d(f^*) + && \longrightarrow \text{Bias error} \Rightarrow \text{Underfitting} \\ & + \ell_D(f^*) && \longrightarrow \text{Bayes risk} \Rightarrow \text{Unavoidable} \end{aligned}$$

Even the best predictor is going to suffer that

$f^*$  is **Bayes Optimal** for  $(D, \ell)$

$$\forall h \quad \ell_D(h) \geq \ell_D(f^*)$$

If  $f^* \notin H$  then  $\ell_D(h^*) > \ell_D(f^*)$

If i pick  $h^*$  I will pick some error because we are not close enough to the risk.

We called this component **bias error**.

Bias error is responsible for underfitting (when training and test are close to each but they are both high :( )

**Variance error** over fitting

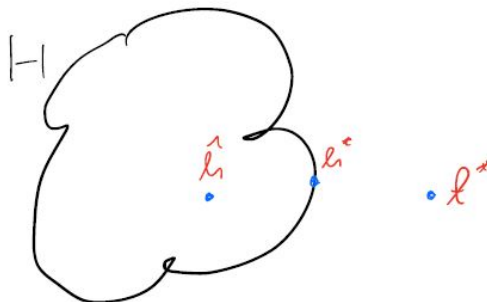


Figure 7.6: Draw of how  $\hat{h}$ ,  $h^*$  and  $f^*$  are represented

Variance is a random quantity and we want to study this. We can always get risk from training error.

## 7.3 Studying overfitting of a ERM

We can bound it with probability.

**I add and subtract trivial training error  $\hat{\ell}_S(h)$**

$$\begin{aligned} \ell_D(\hat{h}) - \ell_D(h^*) &= \ell_D(\hat{h}) - \hat{\ell}_S(h) + \hat{\ell}_S(\hat{h}) - \ell_D(h^*) \leq \\ &\leq \ell_D(\hat{h}) - \hat{\ell}_S(\hat{h}) + \hat{\ell}_S(h^*) - \ell_D(h^*) \leq \\ &\leq |\ell_D(\hat{h}) - \hat{\ell}_S(h)| + |\hat{\ell}_S(h^*) - \ell_D(h^*)| \leq \\ &\leq 2 \cdot \max |\hat{\ell}_S(h) - \ell_D(h)| \end{aligned}$$

(no probability here)

**Any given  $\hat{h}$  minimising  $\hat{\ell}_S(h)$**

Now assume we have a large deviation

$$\text{Assume } \ell_D(\hat{h}) - \ell_D(h^*) > \varepsilon \quad \Rightarrow \quad \max |\hat{\ell}_S(h) - \ell_D(h)| > \frac{\varepsilon}{2}$$

We know  $\ell_D(\hat{h}) - \ell_D(h^*) \leq 2 \cdot \max |\hat{\ell}_S(h) - \ell_D(h)| \Rightarrow$

$$\Rightarrow \exists h \in H \quad |\hat{\ell}_S(h) - \ell_D(h)| > \frac{\varepsilon}{2} \quad \Rightarrow$$

with  $|H| < \infty$

$$\Rightarrow U(|\hat{\ell}_S(h) - \ell_D(h)|) > \frac{\varepsilon}{2}$$

$$\begin{aligned} \mathbb{P}(\ell_D(\hat{h}) - \ell_D(h^*) > \varepsilon) &\leq \mathbb{P}\left(U(|\hat{\ell}_S(h) - \ell_D(h)|) > \frac{\varepsilon}{2}\right) \leq \\ &\leq \sum_{h \in H} \mathbb{P}\left(|\hat{\ell}_S(h) - \ell_D(h)| > \frac{\varepsilon}{2}\right) \leq \sum_{h \in H} 2 \cdot e^{-2\left(\frac{\varepsilon}{2}\right)^2 m} \leq \end{aligned}$$

**Union Bound Chernoff. Hoeffding bound ( $\mathbb{P}(\dots)$ )**

$$\leq 2 \cdot |H| e^{-\frac{\varepsilon^2}{2} m}$$

Solve for  $\varepsilon$   $2 \cdot |H| e^{-\frac{\varepsilon^2}{2} m} = \delta$

$$\text{Solve for } \varepsilon \longrightarrow \varepsilon = \sqrt{\frac{2}{m} \cdot \ln \cdot \frac{2|H|}{\delta}}$$

$$\ell_D(\hat{h}) - \ell_D(h^*) \leq \sqrt{\frac{2}{m} \cdot \ln \cdot \frac{2|H|}{\delta}}$$

With probability at least  $1 - \delta$  with respect to random draw of  $S$ .  
We want  $m \gg \ln|H| \rightarrow$  in order to avoid overfitting

# Lecture 8 - 31-03-2020

$$|H| < \infty \quad \hat{h} = \operatorname{argmin} \hat{\ell}_S(h) \quad h^* = \operatorname{argmin} \ell_D(h)$$

minimise risk

## Bias-Variance decomposition

$$\begin{aligned} \ell_D(\hat{h}_S) &= \ell_D(\hat{h}_S) - \ell_D(h^*) + && \rightarrow \text{Variance error} \Rightarrow \text{Overfitting} \\ &+ \ell_D(h^*) - \ell_D(f^*) + && \rightarrow \text{Bias error} \Rightarrow \text{Underfitting} \\ &+ \ell_D(f^*) && \rightarrow \text{Bayes risk} \Rightarrow \text{Unavoidable} \end{aligned}$$

We state this for all algorithm but we studied for ERM.

$$\ell_D(\hat{h}_S) \leq \ell_D(h^*) + \sqrt{\frac{2}{m} \ln \frac{2|H|}{\delta}} \quad \text{with probability at least } 1 - \delta \text{ over the draw of } S$$

we want **this** to be small when  $m \gg \ln |H|$

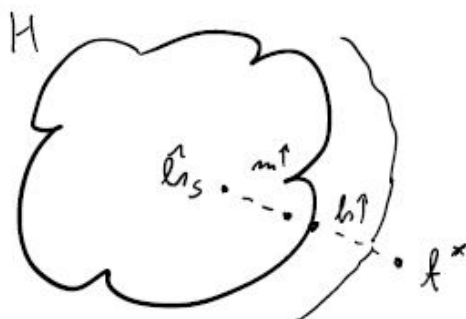


Figure 8.1: Representation of  $\hat{h}$ ,  $h^*$  and  $f^*$

Size of model fix, increase  $m$  (size of sample) when  $m$  is bigger  $\rightarrow$  variance goes down.

When  $|H| \rightarrow$  big model will be closer to optimal

- $m$  grows  $\Rightarrow$  variance error goes down (if not overfitting)
- $|H|$  grows  $\Rightarrow$  bias error goes down (if not underfitting)

ERM with  $|H| < \infty$   
 $A$   $H$  such that  $\forall s A(S) \in H$

We controlled this event:

$$\forall h \in H \quad \hat{\ell}_S(h) - \ell_D(h) \leq \sqrt{\frac{2}{m} \ln \frac{2|H|}{\delta}} \quad \text{with probability at least } 1 - \delta$$

We assure that training error is a good proxy for the true risk.

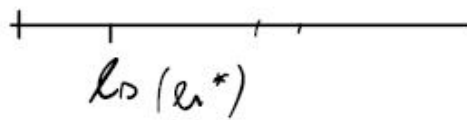


Figure 8.2: Example

**If I do the empirical way (for a specific training set) i get something different.**

## 8.1 The problem of estimating risk in practise

Test error is a good estimate for a risk, provided  $\frac{1}{\sqrt{n}}$  is small

It's usually good to take a small test set.

80/20 big data, if low data better to look at good estimate...=??

Usually small test set is still ok

Typically we are not given the predictor, instead we start from a learning algorithm. It's true that  $A$  has parameter (how many nodes for a tree classifier for example).

In general, if I have parameters then i get a set of different classifier Imaging Algorithm that has parameter  $\theta$ :

$$A \quad \{A_\theta : \theta \in \Theta\} \quad A_\theta(S) = \hat{h}_S$$

$\mathbb{E}[\ell_D(A_\theta(S))]$  **typically averaged over  $S$  of size  $m$  for fixed  $m$**

In general you may also be interest looking at the best choice of parameter for your algorithm: Suppose now that i want to look at mini sing the risk



with the respect to the choice of the parameter

$\mathbb{E}[\min \ell_D(A_\theta(S))]$  i want to choose parameter that minimise risk running algorithm on training set.

I would like to choose best possible value for my parameter  $k$  in the  $k$ -NN and i want to estimate the risk.

$\theta$  can be a set of parameters (so more than 1 parameter), so theta are the **hyper parameters of A**.

In general this parameter are the choices that algorithm that can make: parameter that define a classifier (example nodes and test in the internal nodes and label in the leaves)

**Hyper parameters** are not determined by the training set  $\rightarrow$  chosen before the training set.

I fix  $k$  for  $K_{NN}$  and I choose an upper bound of number of nodes after the training set is given.

So, some parameters are given after training set is given and some parameter fixed before.

Now the idea is that parameter are determined after training set is given, while hyper parameters are given before training set to get then a predictor. I have a family of algorithms, so I choose a hyper parameters to get a one algorithm.

Most algorithm are given as family. We have first decide hyper parameters not determined on the training set.

One way to move is to take you dataset and the split that in 3 parts:

- Training set
- Development (or validation) test
- Test sets

There should not be a leak of information between test and train and development.

We get family of algorithm, train with train set and then use dev set to test the parameter and choice the parameters. Once i found the parameter I re-train the algorithm with a part of train and dev set to being then tested.

**Development set is like a fake test set**  $\rightarrow$  it usefull to choose parameters.

Algorithm steps:

1. Train  $A_\theta$  on training set for each  $\theta \in \Theta$  (**grid search**)

2. Find  $\theta$  such that  $\hat{h} = A_{\hat{\theta}}(S)$  minimise development error
3. Train  $A_{\hat{\theta}}$  on training + development set
4. Test resulting predictor on test set

(There's theory about this but it's difficult and we are not going to do that)

It's heuristic and kinda simple to do. This technics work for every learning algorithm.

One parameter: grid on this

Two parameters ecc..

**It's quadratic!**

**Good learning algorithm should have small number of hyperparameters.**

## 8.2 Cross-validation

Solving an easier problem: suppose you have a dataset and what you do is that you can choose training set and test set.

Algorithm  $A$  with no hyper parameters and we would like to check how good is the predictor: how good can  $A$  be?  $A$  is good if the predictor that generate has low risk.

I split dataset in training and test set.



Figure 8.3: Splitting test and training set

$$h = A(A) \quad \hat{\ell}_{S'}(h) \approx \ell_D(h)$$

I can use Cross-validation! (CV). It helps estimate the risk.

$$\mathbf{CV:} \quad \mathbb{E}[\ell_D(A(S))]$$

I would like to average ....

How do i do this estimate? Super easy, take my data assuming CV as parameter not of the algorithm A, but intrinsic to cross validation.

Parameter k-fold CV typically  $k = 5$  or  $10$ .

What do you do ? If i take a specific training set, i got a ruff estimate of A.

I shouldn't split one but several time.

There are different way but he give us another:

**split data in k folds randomly!**

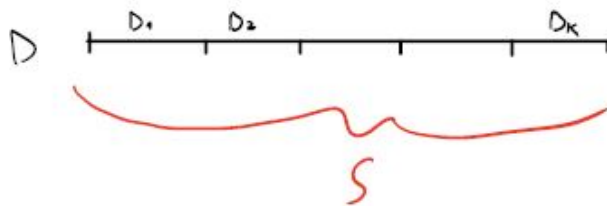


Figure 8.4: K-folds

For each  $k$  we define  $S^{(k)}$  **take out k-st D**

$$S^{(k)} = S \setminus D_k$$

$$S^{(1)} = D_2 \cup D_3 \cup \dots \cup D_k$$

$$S^{(2)} = D_1 \cup D_3 \cup \dots \cup D_k$$

For each  $k = 1, \dots, k$  **folds**  $S^{(k)}$  training part and  $D_k$  test part

$$h_K = A(S^{(k)}) \quad \hat{\ell}_{DK}(h_K) = \frac{k}{m} \sum_{(x,y) \in D_k} \ell(y, h_K(x))$$

Repeat the procedure for  $k = 1 \dots k$  get  $h_1, \dots, h_k$

$$\text{Compute} \quad \frac{1}{k} \cdot \sum_{k=1}^K \hat{\ell}_{DK}(h_k)$$

where  $CV$  is  $\mathbb{E}[\ell_D(A(S))]$  and this is called ..... —MANCA — Estimate

It's used a lot!

You get data from internet, then what to do? I want to try an algorithm, try  $K_{NN}$  so you do Cross validation and will give you the risk.

In some other cases you get a splitted dataset in training and testing set.

You don't use CV since the dataset is already splitted in training and test.

## 8.3 Nested cross validation

The use of CV to solve the hyper parameters choice problem. If you are given test and training set you can solve splitting in training set and dev set. Suppose not given train and test, you can assign arbitrary .....

Now the idea: you give me a way to optimise splitting training set and test set and then avoiding using a ...

**This is what cross validation is doing.**

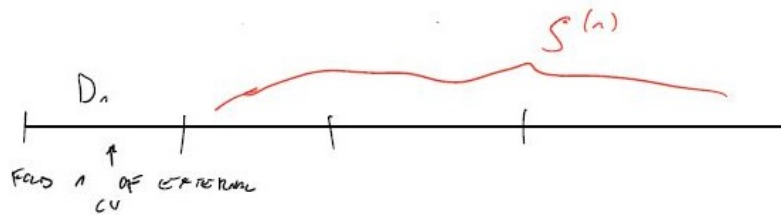


Figure 8.5: Nested Cross Validation

$\{A_\theta : \theta \in \Theta\}$  which i run in each fold?

The idea is **running interval CV for the folds**

I have fold  $D_1$  and  $S(1)$  and i perform external validation, then i run internal CV in  $S(1)$ .

On each training part of the external CV, run a internal CV for each  $A_\theta$ .

When  $\theta \in \text{grid}(\Theta)$ :

- I have a CV-estimate and for each  $A_\theta$  pick the  $\theta$  with best CV-estimate
- Run  $A_{\hat{\theta}}$  on the entire training part of current external fold

**Basically what I'm choosing the best hyper parameter on each fold of the external CV.**

External CV is not testing  $A_{\hat{\theta}}$  for a given  $\hat{\theta} \in \Theta$

I am not measuring a goodness of predictor generate by algorithm for given value of hyper parameters but what I'm estimating is the average risk of the predictor output by learning algorithm when hyper parameters are optimise on the training set. This optimisation on training set is done into a internal CV. To avoid be depend i run and external CV.

Many platform like sklearn allow you to do that in two lines of code. So i can specify the grid, predictor, number of falls for internal and external. It took a bit but that's it in in two lines of code

Cross validation can be done in every order of  $D_1$  or  $D_2$  or  $D_k$ . So doesn't matter the order we start fold  $D_2$  or  $D_1$ .

# Lecture 9 - 06-04-2020

$\hat{h}$  is ERM predictor

$$\ell_D(\hat{h}) \leq \min \ell_D(h) + \sqrt{\frac{2}{m} \ln \frac{2H}{\delta}} \quad \text{with prob. at least } 1 - \delta$$

Now we do it with tree predictors

## 9.1 Tree predictors

$$X = \{0, 1\}^d \longrightarrow \text{Binary classification}$$

$$h : \{0, 1\}^d \longrightarrow \text{Binary classification H1}$$

How big is this class?

Take the size of codomain power the domain  $\longrightarrow |H| = 2^{2^d}$

Can we have a tree predictor that predict every H in this class?

For every  $h : \{0, 1\}^d \longleftarrow \{-1, 1\} \quad \exists T$

We can **build a tree** such that  $h_T = h$

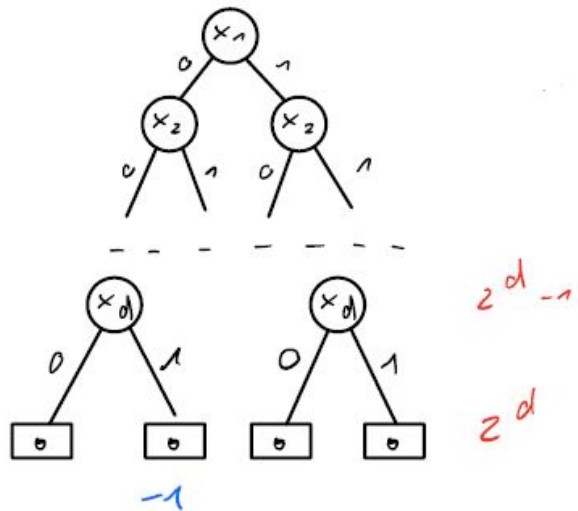


Figure 9.1: Tree building

$$X = (0, 0, 1, \dots, 1) \quad h(x) = -1$$

$x_1, x_2, x_3, \dots, x_d$

I can apply my analysis to this predictors

If I run ERM on  $H$

$$\ell_D(\hat{h}) \leq \min \ell_D(\hat{h}) + \sqrt{\frac{2}{m} 2^d \ln 2 + \ln \frac{2}{\delta}} \quad \rightarrow \ln |H| + \ln \frac{2}{\delta}$$

No sense! What we find about training set that we need?

Worst case of overfitting  $m \gg 2^D = |X| \Rightarrow$  training sample larger

**PROBLEM:** cannot learn from a class too big ( $H$  is too big)

I can control  $H$  just limiting the number of nodes.

$H_N \rightarrow$  tree  $T$  with at most  $N$  nodes,  $N \ll 2^D$

$|H_N| = ?$

$|H_N| = (\# \text{ of trees with } \leq N \text{ nodes}) \times (\# \text{ of test on internal nodes}) \times (\# \text{ labels on leaves})$

$$|H_N| = \otimes \times d^M \times 2^{N-M}$$

$N$  of which  $N - M$  are leaves

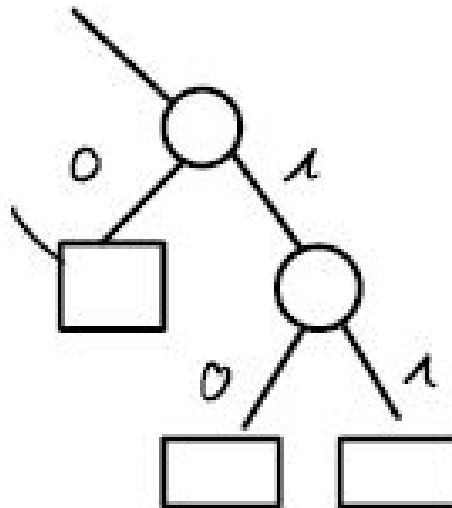


Figure 9.2: Tree with at most  $N$  nodes

$\otimes$  # of binary trees with  $N$  nodes, called **Catalan Number**

### 9.1.1 Catalan Number

\*We are using a binomial \*

$$\frac{1}{N} \binom{2N-2}{N-1} \leq \frac{1}{N} \left( e \frac{(2N-2)}{N-1} \right)^{N-1} = \frac{1}{N} (2e)^{N-1}$$

$$\binom{N}{K} \leq \left( \frac{en}{k} \right)^k \quad \text{from Stirling approximation}$$

Counting the number of tree structure: a binary tree with exactly N nodes. Catalan counts this number.  $\rightarrow$  **but we need a quantity to interpret easily**. So we compute it in another way.

Now we can rearrange everything.

$$|H_N| \leq \frac{1}{N} (2e)^{N-1} H^M 2^{N-M} \leq (2ed)^N$$

$d \geq 2 \leq d^{N-M}$

where **we ignore  $\frac{1}{N}$  since we are going to use the log**

ERM on  $H_N$   $\hat{h}$

$$\ell_D(\hat{h}) \leq \min_{h \in H_N} \ell_D(h) + \sqrt{\frac{2}{m} \left( N \cdot (1 + \ln(2 \cdot d)) + \ln \frac{2}{\delta} \right)}$$

were  $N \cdot (1 + \ln(2 \cdot d)) = \ln(H_N)$

In order to not overfit  $m \gg N \cdot \ln d$

$N \cdot \ln d \ll 2^d$  for reasonable value of N

We grow the tree and a some point we stop.

$$\ell_D(h) \leq \hat{\ell}_S(h) + \varepsilon \quad \forall h \in H_N \quad \text{with probability at least } 1 - \delta$$

**remove N in  $H_N$  and include h on  $\varepsilon$**

we remove the N index in  $H_N$  adding h on  $\varepsilon$

$$\ell_D(h) \leq \hat{\ell}_S(h) + \varepsilon_h \quad \forall h \in H_N$$

$$W : H \rightarrow [0, 1] \quad \sum_{h \in H} w(h) \leq 1$$

**How to use this to control over risk?**

$$\mathbb{P} \left( \exists h \in H : |\hat{\ell}_S(h) - \ell_D(h)| > \varepsilon_h \right) \leq$$

where  $\hat{\ell}_S$  is the prob my training set cases is true

$$\begin{aligned} &\leq \sum_{h \in H} \mathbb{P} \left( |\hat{\ell}_S(h) - \ell_D(h)| > \varepsilon_h \right) \leq \sum_{h \in H} 2e^{-2m\varepsilon_h^2} \leq \\ &\leq \delta \quad \rightarrow \text{since } w(h) \text{ sum to } 1 \left( \sum_{h \in H} \right) \end{aligned}$$

I want to choose  $2e^{-2m\varepsilon_h^2} = \delta w(h)$

$$2e^{-2m\varepsilon_h^2} = \delta w(h) \quad \Leftrightarrow \quad \text{--- MANCA PARTEEEE ---}$$

therefore:

$$\ell_D(h) \leq \hat{\ell}_S(h) + \sqrt{\frac{1}{2m} \cdot \left( \ln \frac{1}{w(h)} + \ln \frac{2}{\delta} \right)} \quad \text{w. p. at least } 1 - \delta \quad \forall h \in H$$

Now, instead of using ERM we use

$$\hat{h} = \arg \min_{h \in H} \left( \hat{\ell}_S(h) + \sqrt{\frac{1}{2m} \cdot \left( \ln \frac{1}{w(h)} + \ln \frac{2}{\delta} \right)} \right)$$

where  $\sqrt{\dots}$  term is the penalisation term

Since our class is very large we add this part in order to avoid overfitting. Instead of minimising training error alone i minimise training error + penalisation error.

In order to pick  $w(h)$  we are going to use **coding theory**

The idea is I have my trees and i want to encode all tree predictors in  $H$  using strings of bits.

$\sigma : H \rightarrow \{0, 1\}^*$  **coding function for trees**  
 $\forall h, h' \in H \quad \sigma(h)$  not a prefix of  $\sigma(h')$   
 $h \neq h' \quad$  where  $\sigma(h)$  and  $\sigma(h')$  are **string of bits**

$\sigma$  is called **instantaneous coding function**

Instantaneous coding function has a property called **kraft inequality**

$$\sum_{h \in H} 2^{-|\sigma(h)|} \leq 1 \quad w(h) = 2^{-|\sigma(h)|}$$



I can design  $\sigma : H \rightarrow \{0, 1\}^*$  instantaneous  $|\sigma(h)|$

$$\ln |H_N| = O(N \cdot \ln d)$$

**number of bits i need = number of node in  $h$**

Even if i insist in instantaneous i do not lose ... - MANCA PARTE -

$$|\sigma(h)| = O(N \cdot \ln d)$$

Using this  $\sigma$  and  $w(h) = 2^{-|\sigma(h)|}$

$$\ell_D(h) \leq \hat{\ell}_S(h) + \sqrt{\frac{1}{2m} \cdot \left( c \cdot N \cdot \ln d + \ln \frac{2}{\delta} \right)} \quad w. p. \text{ at least } 1 - \delta$$

where  $c$  is a constant

$$\hat{h} = \arg \min_{h \in H} \left( \hat{\ell}_S(h) + \sqrt{\frac{1}{2m} \cdot \left( c \cdot N \cdot \ln d + \ln \frac{2}{\delta} \right)} \right)$$

where  $m \gg N \cdot h \cdot \ln d$

If training set size is very small then you should not run this algorithm.

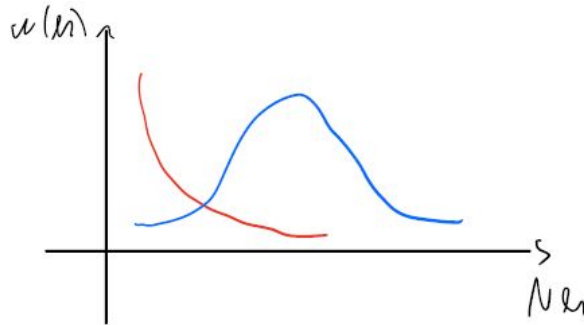


Figure 9.3: Algorithm for tree predictors

This blue curve is an alternative example. We can use Information criterion.

As I increase the number of nodes,  $N_h$  decrease so fast. You should take a smaller tree because it gives you a better bound. It's a principle known as Occam Razor ( if I have two tree with the same error, if one is smaller than the other than i should pick this one).

Having  $N^*$

# Lecture 10 - 07-04-2020

## 10.1 TO BE DEFINE

## 10.2 MANCANO 20 MINUTI DI LEZIONE

$$\mathbb{E}[z] = \mathbb{E}[\mathbb{E}[z|x]] \quad \longrightarrow \quad \mathbb{E}[Z|X=x]$$

$$\mathbb{E}[X] = \sum_{t=1}^m \mathbb{E}[x \cdot \Pi(A_t)] \quad A_1, \dots, A_m \text{ portion of sample law of total probability}$$

$$\begin{aligned} x \in \mathbb{R}^d \quad \mathbb{P}(Y_{\Pi(s,x)} = 1) &= \mathbb{E}[\Pi Y_{\Pi(s,x)} = 1] = \quad \text{Law of total probability} \\ &= \sum_{t=1}^m \mathbb{E}(\Pi\{Y_t = 1\} \cdot \Pi \cdot \{\Pi(s,x) = t\}) = \\ &= \sum_{t=1}^m \mathbb{E}[\mathbb{E}[\Pi\{Y_t = 1\} \cdot \Pi \cdot \{\Pi(s,x) = t\} | X_t]] = \end{aligned}$$

given the fact that  $Y_t \sim \eta(X_t) \Rightarrow$  give me probability

$$\begin{aligned} Y_t = 1 \text{ and } \Pi(s,x) = t \text{ are independent given } X_Y \quad (e.g. \mathbb{E}[Z|X] = \mathbb{E}[x] \cdot \mathbb{E}[z]) \\ = \sum_{t=1}^m \mathbb{E}[\mathbb{E}[\Pi\{Y_t = 1\} | X_t] \cdot \mathbb{E}[\Pi(s,x) = t | X_t]] = \sum_{t=1}^m \mathbb{E}[\eta(X_t) \cdot \Pi \cdot \{\Pi(s,x) = t\}] = \\ = \mathbb{E}[\eta(X_{\Pi(s,x)})] \end{aligned}$$

$$\mathbb{P}(Y_{\Pi(s,x)} | X = x) = \mathbb{E}[\eta(X_{\Pi(s,x)})]$$

$$\begin{aligned} \mathbb{P}(Y_{\Pi(s,x)} = 1, y = -1) &= \mathbb{E}[\Pi\{Y_{\Pi(s,x)} = 1\} \cdot \Pi\{Y = -1|X\}] = \\ &= \mathbb{E}[\Pi\{Y_{\Pi(s,x)} = 1\} \cdot \Pi\{y = -1\}] = \mathbb{E}[\mathbb{E}[\Pi\{Y_{\Pi(s,x)} = 1\} \cdot \Pi\{y = -1|X\}]] = \end{aligned}$$

by independence i can split them

$$Y_{\Pi(s,x)} = 1 \quad y = -1 \quad \text{which is } 1 - \eta(x) \quad \text{when } X = x$$

$$= \mathbb{E} [\mathbb{E} [\Pi\{Y_{\Pi}(s, x)\} = 1 | X] \cdot \mathbb{E} [\Pi\{y = -1\} | X]] = \mathbb{E} [\eta_{\Pi(s, x)} \cdot (1 - \eta(x))] =$$

similarly:

$$\mathbb{P}(Y_{\Pi(s, x)} = -1, y = 1) = \mathbb{E} [(1 - \eta_{\Pi(s, x)}) \cdot \eta(x)]$$

$$\begin{aligned} \mathbb{E} [\ell_D(\hat{h}_s)] &= \mathbb{P}(Y_{\Pi(s, x)} \neq y) = \mathbb{P}(Y_{\Pi(s, x)} = 1, y = -1) + \mathbb{P}(Y_{\Pi(s, x)} = -1, y = 1) \\ &= \mathbb{E} [\eta_{\Pi(s, x)} \cdot (1 - \eta(x))] + \mathbb{E} [(1 - \eta_{\Pi(s, x)}) \cdot \eta(x)] \end{aligned}$$

**Make assumptions on  $D_x$  and  $\eta$ :**

1.  $\forall X$  drawn from  $D_x$   $\max |X_t| \leq 1$   
Feature values are bounded in  $[-1, 1]$   
all my points belong to this:

$$X = [-1, 1]^d$$

2.  $\eta$  is such that  $\exists c < \infty$  :

$$\eta(x) - \eta(x') \leq c \cdot \|X - x'\| \quad \forall x, x' \in X$$

It means that  $\eta$  is **Lipschitz** in  $X$   $c < \infty \Leftrightarrow \eta$  is continuous

**using two facts:**

$$\eta(x') \leq \eta(x) + c \|X - x'\| \quad \longrightarrow \quad \text{euclidean distance}$$

$$1 - \eta(x') \leq 1 - \eta(x) + c \|X - x'\|$$

$$X' = X_{\Pi(s, x)}$$

$$\begin{aligned} &\eta(X) \cdot (1 - \eta(x')) + (1 - \eta(x)) \cdot \eta(x') \leq \\ &\leq \eta(x) \cdot ((1 - \eta(x)) + \eta(x) \cdot c \|X - x'\|) + (1 - \eta(x)) \cdot c \|X - x'\| = \\ &= 2 \cdot \eta(x) \cdot (1 - \eta(x)) + c \|X - x'\| \end{aligned}$$

$$\mathbb{E} [\ell_d \cdot (\hat{h}_s)] \leq 2 \cdot \mathbb{E} [\eta(x) - (1 - \eta(x))] + c \cdot (E) [\|X - x_{\Pi(s, x)}\|]$$

where  $\leq$  mean at most

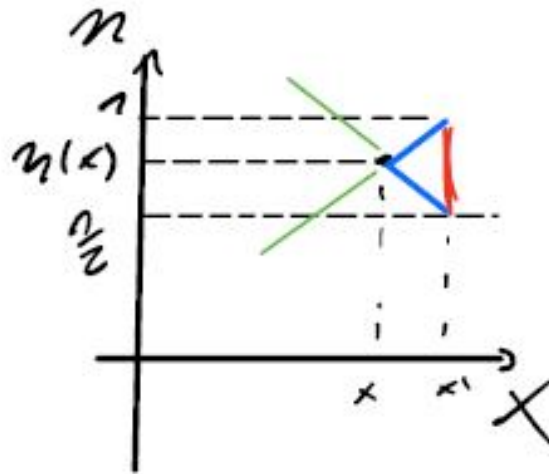


Figure 10.1: Point (2) - where  $y = cx + q$   $y = -cx + q$

### 10.3 Compare risk for zero-one loss

$$\mathbb{E}[\min\{\eta(x), 1 - \eta(x)\}] = \ell_D(f^*)$$

$$\eta(x) \cdot (1 - \eta(x)) \leq \min\{\eta(x), 1 - \eta(x)\} \quad \forall x$$

$$\mathbb{E}[\eta(x) \cdot (1 - \eta(x))] \leq \ell_D(f^*)$$

$$\mathbb{E}[\ell_d(\hat{l}_s)] \leq 2 \cdot \ell_D(f^*) + c \cdot \mathbb{E}[\|X - X_{\Pi(s,x)}\|] \quad \eta(x) \in \{0, 1\}$$

Depends on dimension: **curse of dimensionality**

$\ell_d(f^*) = 0 \iff \min\{\eta(x), 1 - \eta(x)\} = 0$  with probability = 1  
to be true  $\eta(x) \in \{0, 1\}$

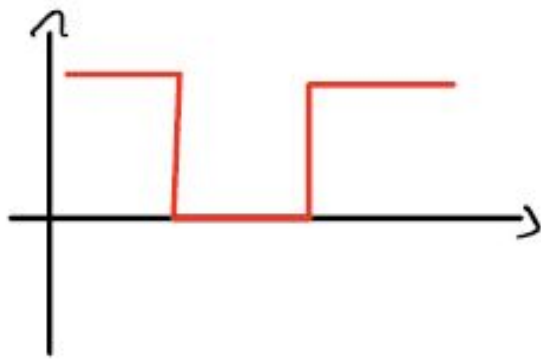


Figure 10.2: Point

# Lecture 11 - 20-04-2020

## 11.1 Analysis of $K_{NN}$

$$\mathbb{E} \left[ \ell_D(\hat{\ell}_s) \right] \leq 2 \cdot \ell_D(f^*) + c \cdot \mathbb{E} \left[ \|X - x_{\Pi(s,x)}\| \right]$$

At which rate this thing goes down? If number of dimension goes up then a lot of point are far away from  $X$ .

So this quantity must depend on the space in which  $X$  live.

Some dependence on number of depends and increasaing number of traning points close to  $X$

This expection is fucniton of random variable  $X$  and  $X_{\pi(s,x)}$

We are going to use the assumption that:

$$|X_t| \leq 1 \quad \forall \text{ cordinates } i = 1, \dots, d$$

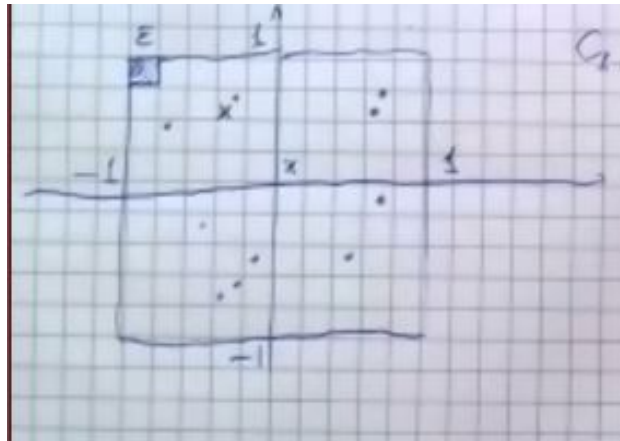


Figure 11.1: Example of domain of  $K_{NN}$

Hyper box in bydimension. All point live in this box and we exploit that. Look at the little suare in which is divided and we assume that we are dividing the box in small boxes of size  $\varepsilon$ . Now the training points will be a strinle of point distributed in the big square.

Our training points are distributed in the box (this is our  $S$ ).

Now we added a point  $x$  and given this two things can happned: falls in the square with training points or in a square without training points.

What is going to be the distance  $X_{\pi(s,x)}$  in this two cases?

We have  $c_1$  up to  $c_r$  How big is this when we have this two cases? (We

looking at specific choices of  $x$  and  $s$ )

$$\|X - X_{s,x}\| \leq \begin{cases} \varepsilon\sqrt{d} & C_i \cup S \neq 0 \\ \sqrt{d} & C_i \cup S = 0 \end{cases}$$

were  $X \in C_i$

We have to multiply by the length of the cube. Will be  $\varepsilon\sqrt{d}$

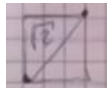


Figure 11.2: Diagonal length

If things go badly can be very far away like the length of the domain. Length is 2 and diagonal is  $\sqrt{d}$

if close they are going to be  $\varepsilon$  close or far as domain.

We can split that the expression inside the expectation according to the two cases.

$$\begin{aligned} & \mathbb{E} [\|X - X_{\Pi(s,x)}\|] \leq \\ & \mathbb{E} \left[ \varepsilon \cdot \sqrt{d} \cdot \sum_{i=1}^r I\{X \in C_i\} \cdot I\{C_i \cap S \neq 0\} \right] + 2 \cdot \sqrt{d} \cdot \sum_{i=1}^r I\{X \in C_i\} \cdot I\{C_i \cap S = 0\} = \\ & = \varepsilon \cdot \sqrt{d} \cdot \mathbb{E} \left[ \sum_{i=1}^r I\{X \in C_i\} \cdot I\{C_i \cap S \neq 0\} \right] + 2 \cdot \sqrt{d} \cdot \sum_{i=1}^r \mathbb{E} [I\{X \in C_i\} \cdot I\{C_i \cap S = 0\}] \leq \end{aligned}$$

I don't care about this one  $\sum_{i=1}^r I\{X \in C_i\} \cdot I\{C_i \cap S \neq 0\}$

Can be either 0 or 1 (if for some  $i$ ,  $X$  belong to some  $C_i$ )

So at most 1 the expectation

$$\leq \varepsilon \cdot \sqrt{d} + \square$$

We can bound this square. Are the event  $I$  in the summation of the term after  $+$ . If they are independent the product will be the product of the two expectation. If I fix the cube.  $X$  and  $S$  are independent.

Now the two events are independent

**$X \in C_i$  is independent of  $C_i \cap S \neq 0$**

$$\mathbb{E} [I\{X \in C_i\} \cdot I\{C_i \cap S \neq 0\}] = \mathbb{E} [I\{X \in C_i\}] \cdot \mathbb{E} [I\{C_i\}]$$



MANCAAAAAAAAA 9.26

$$\mathbb{P}(C_i \cap S) = (1 - \mathbb{P}(X \in C_1))^m \leq \exp(-m \cdot \mathbb{P}(x \in C_1))$$

The probability of the point fall there and will be the probability of falling in the cube.

Probability of Xs to fall in the cube with a m (samples?)

Now use inequality  $(1 - p)^m \in e^{-pm} \rightarrow 1 + x \leq e^x$

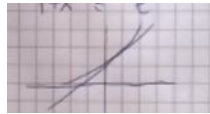


Figure 11.3: Shape of the function

$$\sum_{t=1}^r \mathbb{E}[\mathbb{P}(X \in C_1) \cdot \mathbb{P}(C_1 \cap S \neq)] \leq \sum_{i=1}^r p_i \cdot e^{-mp_i} \leq$$

given that  $p_i = \mathbb{P}(X \in C_i)$  I can upper bound this

$$\leq \sum_{t=1}^r \left( \max_{0 \leq p \leq 1} p e^{-mp} \right) \leq r \max_{0 \leq p \leq 1} p e^{-mp} =$$

where  $p e^{-mp}$  is  $F(p)$  it is concave function so i'm going to take first order derivative to maximise it.

$$F'(p) = 0 \Leftrightarrow p = \frac{1}{m} \quad \text{check!}$$

$$F''(p) \leq 0$$

Check this two condition!

$$= \frac{r}{em}$$

Now get expectation

$$\mathbb{E}[\|X - X_{\Pi(s,x)}\|] \leq \varepsilon \cdot \sqrt{d} + \left(2 \cdot \sqrt{d}\right) \frac{r}{em} =$$

I have  $\left(\frac{2}{\varepsilon}\right)^2$  squares. This bring  $\varepsilon$  in the game

$$\varepsilon \cdot \sqrt{d} + \left(2 \cdot \sqrt{d}\right) \frac{1}{em} \cdot \left(\frac{2}{\varepsilon}\right)^d =$$

$$= \sqrt{d} \left( \varepsilon + \frac{2}{em} \cdot \left( \frac{2}{\varepsilon} \right)^d \right)$$

**HE MISS THE "c" costant from the start** we can choose  $\varepsilon$  to take them balanced  
 set  $\varepsilon = 2m^{-\frac{1}{d+1}}$

$$\left( \varepsilon + \frac{2}{em} \right) \cdot \left( \frac{2}{\varepsilon} \right)^d \leq 4m^{-\frac{1}{d+1}} =$$

$$\mathbb{E} \left[ \ell_d(\hat{h}_s) \right] \leq 2\ell_d(f^*) + 4 \cdot c \cdot \sqrt{d} \cdot m^{-\frac{1}{d+1}}$$

We have that:

$$\text{if } m \rightarrow \infty \quad \ell_D(f^*) \leq \mathbb{E} \left[ \ell_D(\hat{h}_s) \right] \leq 2\ell_D(f^*)$$

I want this smaller than twice risk + some small quantity

$$\mathbb{E} \left[ \ell_d(\hat{h}_s) \right] \leq 2\ell_D(f^*) + \varepsilon$$

How big  $m$  ?

Ignore this part since very small ( $4 \cdot c \cdot \sqrt{d}$ )

$$m^{-\frac{1}{d+1}} \leq \varepsilon \Leftrightarrow m \geq \left( \frac{1}{\varepsilon} \right)^d + 1$$

So 1-NN require a training set size exponential "accuracy"  $1 - \varepsilon$

We show that 1-NN can approach twice based risk  $2 \cdot \ell_D(f^*)$   
 but it takes a training set exponential in  $d$ .

### 11.1.1 Study of $K_{NN}$

Maybe we can use the  $K_{NN}$ .

$$\mathbb{E} \left[ \ell_D(\hat{h}_s) \right] \leq \left( 1 + \sqrt{\frac{8}{k}} \right) \ell_D(f^*) + O \left( k m^{-\frac{1}{d+1}} \right)$$

So is not exponential here.

**Learning algorithm A is consistent for a certain loss  $\ell$**

If  $\forall D$ (distribution) of data we have that  $A(S_m)$  predictor output by A

Now have the risk of that in  $\ell_D(A(S_m))$  and we look at the expectation

$$\mathbb{E} [\ell_D(A(S_m))]$$

If we give a training set size large (  $\lim_{m \rightarrow \infty} \mathbb{E} [\ell_D(A(S_m))] = \ell_D(f^*)$  ) risk will converge in based risk.

$K_{NN}$  where  $K = K_m$  (is a function of training set size).  $K, \rightarrow \infty$  as  $m \rightarrow \infty$ . Only way  $K$  goes to infinity is sublinearly of training set size. (infinity but so as quickly as  $m$   $K_m = O(m)$ )

For instance  $K_m = \sqrt{m}$

Then:

$$\lim_{m \rightarrow \infty} \mathbb{E} [\ell_D(A'(S_m))] = \ell_D(f^*) \quad \text{where } A' \text{ is } K_m\text{-}NN$$

Increasing the size we will converge to this base risk for any distribution and that's nice.

### 11.1.2 study of trees

Algorithm that grow tree classifiers can also be made consistent provided two condition:

- The tree keeps growing
- A non-vanishing fraction of training example is routed to each leaf

Tree has to keep growing but not so fast.

Second point is: suppose you have a certain number of leaves and you can look at the fraction. Each leaf  $\ell$  gets  $N_\ell$  examples. You want that this fraction at any point of time is not going to 0. The fraction of point every leaf receive a split we are reducing the smallest number of examples.

Example keep growing and leaves too and we want that  $\frac{N_\ell}{manca}$  this not going to 0. . since not showed the formula.

Given  $A$ , how do I know wheter  $A$  could be consistent?

$$H_A \equiv \{ h : \exists S A(S) = h \}$$

$S$  can be any size. If  $A$  is *ERM* then  $H_A = H$ , so where *ERM* minimise it. If  $\exists f^* : X \rightarrow$  such that  $f^* \notin H_A$  and  $\exists D$  such that  $f^*$  is Bayes optimal for some distribution  $D$ .

This cannot be consistent because distribution will not be able to generate the Bayes optimal predictor. Maybe is there another predictor  $f$  which is

not equal to  $f^*$  risk.

What's the intuition?

Every time  $A$  is such that  $H_A$  is "restricted" in some sense, then  $A$  cannot be consistent. (e.g *ERM*).

Another way of restricting? Could be tree classifiers with at most  $N$  nodes (bound number of nodes).

How do i know  $N$  is enough to approximate well  $f^*$ . I want to converge the risk of  $f^*$ .

We can introduce a class of algorithm potentially consistent in which space predictor is not restricted.

## 11.2 Non-parametric Algorithms

When they are potentially consistent.

What does it mean?

Non-parametric algorithm have the potential of being consistent and do we know if algorithm is parametric or not?

$A$  is non-parametric if:

- the description of  $A(S_m)$  grows with  $m$

Your predictor is a function and let's assume i can store in any variable a real number with arbitrary precision.

**Any algorithm with bias is inconsistent. So ability to converge to base risk is this.**

How do i know if i have bias or not? this is where non parametric algorithm came.

Let's consider  $K_{NN}$ , how i can describe it? I have to remember distance is made by training points and if i give you more  $S$  the  $m$  will increase. So this is parametric.

More training set for tree, then will grow more, even more larger will be ever growing more.

Any algorithm as a give training points is no parametric, while growing with parametric will stop a some point.

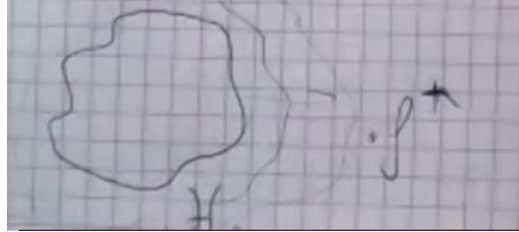


Figure 11.4: Parametric and non parametric growing as training set getting larger

If algorithm is more parametric as i give training points

If a certain point stop growing,  $f^*$  will be out and i will grow more.

If algorithm is able to generate — MANCA — Then the algorithm is non-parametric and can be potentially consistent and include  $f^*$  as it grows.

If set of predictor stops because I'm not enlarging my set of predictor since description of algorithm will not depend on training size at some point → to be consistent.

If bias vanishes as i increase the S, then i can be consistent. I generating predictor that description depends on how much points i give them.

Parametric is not precise as consistency.

One class of algorithm that has consistency has a predictor size growing with S growing.

Definition of non parametric is more fuzzy, consistency is precise (we demonstrate that mathematically).

### 11.2.1 Example of parametric algorithms

Neural network is parametric since i give structure of the network. If i give S small or big S my structure will be the same (will fit better on the training points).

Other example are algorithm with linear classifier in which number of parameter are just the idmension of the space.

# Lecture 12 - 21-04-2020

## 12.1 Non parametrics algorithms

We talk about **consistency**: as the training size grows unbounded the expected risk of algorithms converge to Bayes Risk.

Now we talk about **non parametric algorithm**: the structure of the model is determined by the data.

Structure of the model is fixed, like the structure of a Neural Network but in non parametric algorithm will change structure of the model as the data grows ( $K_{NN}$  and tree predictor).

If I live the tree grow unboundedly then we get a non parametric tree, but if we bound the grows then we get a parametric one.

The converge rate of Bayes Risk (in this case doubled) was small. Converge of 1-NN to  $2\ell_D(f^*)$  is  $m^{-\frac{1}{d+1}}$  so we need an esponential in the dimension. And we need this is under Lips assumption of  $\eta$ .

It's possible to converge to Bayes Risk and it's called **No free lunch**.

### 12.1.1 Theorem: No free lunch

Let a sequeence of number  $a_1, a_2 \dots \in \mathbb{R}$  such that they converge to 0.

Also  $\frac{1}{22222222} \geq a_1 \geq a_2 \geq \dots \forall A$  for binary classification  $\exists D$  s. t.

$\ell_D(f^*) = 0$  (zero-one loss) so Bayes risk is zero and  $\mathbb{E}[\ell_D(A(S_M))] \geq a_m \quad \forall m \geq 1$

Any Bayes Optimal you should be prepared to do so on long period of time. This means that:

- For specific data distribution  $D$ , then  $A$  may converge fast to Bayes Risk.
- If  $\eta$  is Lipschitz then it is continous. This mean that we perturb the input by the output doesn't change too much.
- If Bayes Risk is 0 ( $\ell_D(f^*) = 0$ ) function will be discontinous

This result typically people think twice for using consistent algorithm because

I have Bayes risk and some non consitent algorithm that will converge to some value ( $\ell_D(\hat{h}^*)$ ). Maybe i have Bayes risk and the convergence takes a

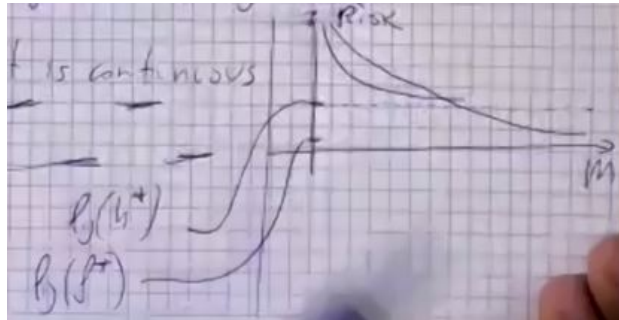


Figure 12.1: Tree building

lot on increasing data points. Before converging was better non parametric (?..)

Picture for binary classification, (similar for other losses)

- Under no assumption on  $\eta$ , the typical "parametric" converge rate to risk of best model in  $H$  (including ERM) is  $m^{-\frac{1}{2}}$ . (Bias error may be high)
- Under no assumption on  $\eta$  there is no guaranteed convergence to Bayes Risk (in general) and this is **no-free-lunch** that guaranteed me no convergence rate.
- Under Lipschitz assumption on  $\eta$  the typical non parametric convergence to Bayes Risk is  $m^{-\frac{1}{d+1}}$ . This is exponentially worst than the parametric convergence rate.

The exponential dependence on  $d$  is called **Curse of dimensionality**.

But if I assume small number of dimension  $\rightarrow K_{NN}$  is ok if  $d$  is small (and  $\eta$  is "easy")

If you have a non parametric algorithm (no Bayes error but may have exponentially infinity training error). I want them to be balanced and avoid bias and variance. We need to introduce a bit of bias in controlled way.

Inserting bias to reducing variance error. So we sacrifice a bit to get a better variance error.

It could be good to inject bias in order to reduce the variance error. In practise instead of having 0 training error i want to have a larger training error and hope to reduce overfitting sacrificing a bit in the training error.

I can increase bias in different technics: one is the unsamle methods.

## 12.2 Highly Parametric Learning Algorithm

### 12.2.1 Linear Predictors

Our domain is Euclidean space (so we have points of numbers).

$$X \text{ is } \mathbb{R}^d \quad x = (x_1, \dots, x_d)$$

A linear predictor will be a linear function of the data points.

$$h : \mathbb{R}^d \rightarrow Y \quad h(x) = f(w^T x) \quad w \in \mathbb{R}^d$$

$$f : \mathbb{R} \rightarrow Y$$

And this is the dot product that is

$$w^T x = \sum_{i=1}^d w_i x_i = \|w\| \|x\| \cos \Theta$$

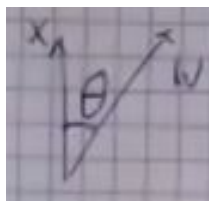


Figure 12.2: Dot product

Suppose we look a regression with square loss.

$$Y = \mathbb{R} \quad h(x) = w^T x \quad w \in \mathbb{R}^d$$

$$f^*(x) = \mathbb{E}[Y|X = x]$$

Binary classification with zero-one loss  $Y = \{-1, 1\}$  We cannot use this since is not a real number but i can do:

$$h(x) = \text{sgn}(w^T x) \quad \text{sgn}(x) = \begin{cases} +1 & \text{if } z > 0 \\ -1 & \text{if } z \leq 0 \end{cases}$$

where  $\text{sgn}$  is a sign function. Linear classifier.

$\|X\| \cos \Theta$  is the length of the projection of  $x$  onto  $w$

Now let's look at this set:

$$\{x \in \mathbb{R}^d : w^T x = c\}$$



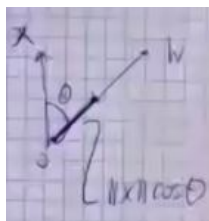


Figure 12.3: Dot product

This is a hyperplane.

$$\|w\| \|x\| \cos \Theta = c \quad \|x\| \cos \Theta = \frac{c}{\|w\|}$$

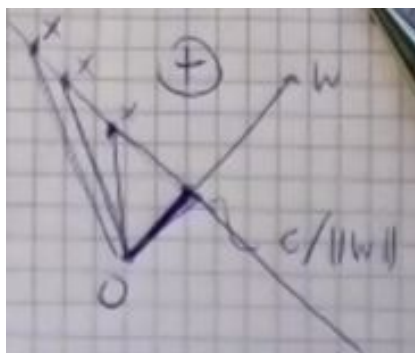


Figure 12.4: Hyperplane

So  $(w, c)$  describe an hyperplane.

We can do binary classification using the hyperplane. Any points that lives in the positive half space and the negative. So the hyperplane is splitting in halves.  $H \equiv \{x \in \mathbb{R}^d : w^T x = c\}$

$$H^+ \equiv \{x \in \mathbb{R}^d : w^T x > c\} \quad \text{positive } h_s$$

$$H^- \equiv \{x \in \mathbb{R}^d : w^T x \leq c\} \quad \text{negative } h_s$$

$$h(x) = \begin{cases} +1 & \text{if } x \in H^+ \\ -1 & \text{if } x \notin H^+ \end{cases} \quad h(x) = \text{sgn}(w^T x - c)$$

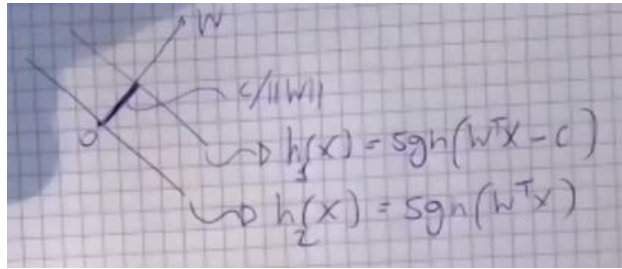


Figure 12.5: Hyperplane

$h_1$  is non-homogenous linear classifier.  
 $h_2$  is homogenous linear classifier. Any homogenous classifier is equivalent to

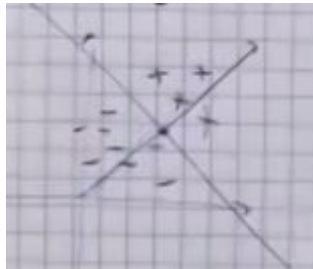


Figure 12.6: Hyperplane

this:

$$\{x \in \mathbb{R}^d : X = c\} \text{ is equivalent to } \{x : \mathbb{R}^{d+1} : \nu^T x = 0\}$$

$$\nu = (w_1, \dots, w_d, -c) \quad x' = (x_1, \dots, x_d, 1)$$

So we added a dimension.

$$w^T x = c \Leftrightarrow \nu^T x' = 0$$

$$\sum_i w_i x_i = c \Leftrightarrow \sum_i w_i x_i - c = 0$$

**Rule:**

When you learn predictor just add an extra feature to your data points, set it to 1 and forget about non-homogenous stuff.

One dimensional example

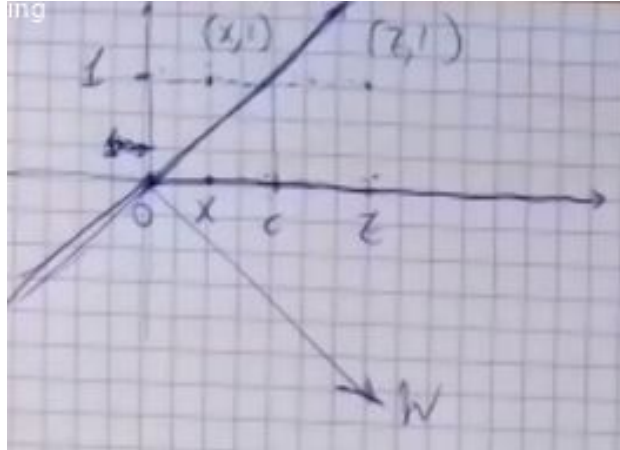


Figure 12.7: Example of one dimensional hyperplane

I have negative (left of  $(x, 1)$ ) and positive point (left of  $(z, 1)$ ) classified

Now i want to learn linear classifier. How can i do it?

$$H_d = \{ h : \exists w \in \mathbb{R}^d h(x) = \text{sgn}(w^T x) \}$$

Parametric!

We expect high bias a low variance.

$$\begin{aligned} \text{ERM} \quad \hat{h}_S &= \arg \min_{h \in H_d} \frac{1}{m} \cdot \sum_{t=1}^m I\{h(x_t) \neq y_t\} = \\ &= \arg \min_{w \in \mathbb{R}^d} \frac{1}{m} \cdot \sum_{t=1}^m I\{y_t w^T x_t \leq 0\} \end{aligned}$$

A bad optimisation problem!

**FACT:**

It is unlikely to find an algorithm that solves ERM for  $H_d$  and zero-one loss efficiently.

**NP completeness problems!**

It's very unlikely to solve this problem.

This problem is called **MinDisagreement**

### 12.2.2 MinDisagreement

Instance:  $(x_1, y_1) \dots (x_m, y_m) \in \{0, 1\}^d \times \{-1, 1\}$ ,  $k \in \mathbb{N}$

Question: Is there  $w \in \mathbb{R}^d$

s.t.  $y_t w^T x_t \leq 0$  for at most  $k$  indices  $t \in \{1, \dots, m\}$

This is NP-complete!

# Lecture 13 - 27-04-2020

## 13.1 Linear prediction

We had  $ERM \hat{h}$

$$S = \{(x_1, y_1) \dots (x_n, y_n)\} \quad x_t \in \mathbb{R}^d \quad y_t \in \{-1, +1\} \quad \ell_t(w) = I\{y_t w^T x_t \leq 0\}$$

$$\hat{h}_S = \arg \min_{h \in H_D} \frac{1}{m} \sum_{t=1}^m I\{y_t w^T x_t \leq 0\}$$

The associated decision problem is a NP problem so cannot be computed efficiently unless  $P \equiv NP$

Maybe we can approximate it, so a good solution that goes close to minimise error.

This is called MinDisOpt

### 13.1.1 MinDisOpt

Instance:  $(x_1, y_1) \dots (x_n, y_n) \in \{0, 1\}^d \times \{0, 1\}$

Solution:

$$w \in Q^D \text{ minimising the number of indices } t = 1, \dots, m \text{ s.t. } h_t w^T x_t \leq 0$$

$Opt(S)$  is the smallest number of misclassified example in  $S$  by any linear classifier in  $H_D$

where  $\frac{Opt(S)}{m}$  is training error of  $ERM$

**Theorem** : if  $P \not\equiv NP$ , then  $\forall c > 0$  there are no polytime algorithms (with r. t. the input size  $\Theta(m_d)$ ) that approximately solve every instance  $S$  of MinDisOpt with a number of mistakes bounded by  $C \cdot Opt(S)$ .

If I am able to approximate it correctly this approximation will grow with the size of the dataset.

$$\forall A \text{ (polytime) and } \forall C \quad \exists S \quad \hat{\ell}_S(A(S)) \geq c \cdot \hat{\ell}_S(\hat{h}_S) \text{ (where } \hat{h}_S \text{ is } ERM)$$

$$Opt(S) = \hat{\ell}_S(\hat{h}_S)$$

This is not related with free lunch theorem (information we need to get base

error for some learning problem). Free lunch: we need arbitrarily information to get such error. Here is we need a lot of computation to approximate the *ERM*.

Assume  $Opt(S) = 0$  *ERM* has zero training error on  $S$   
 $\exists U \in \mathbb{R}^d$  s.t.  $\forall t = 1, \dots, m \quad y_t U^T x_t > 0$       **$S$  is linearly separable**

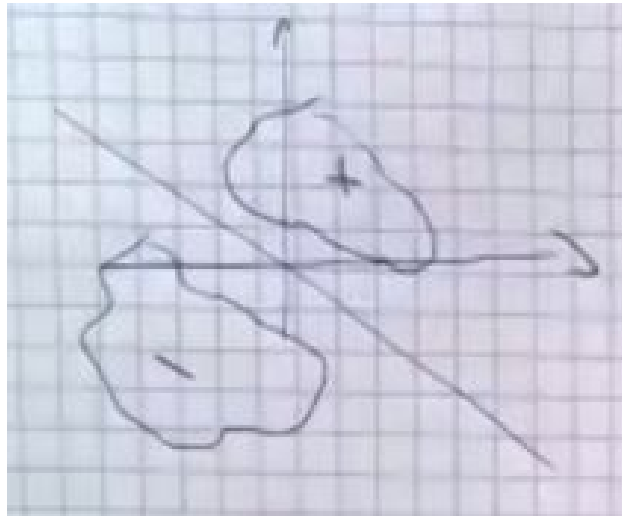


Figure 13.1: Tree building

We can look at the min

$$\min_{t=1, \dots, m} y_t U^T x_t = \gamma(U) > 0 \quad \text{We called this margin of } U \text{ on } (x_t, y_t)$$

We called in this way since  $\frac{\gamma(U)}{\|U\|} = \min_t y_t \|x_t\| \cos(\Theta)$

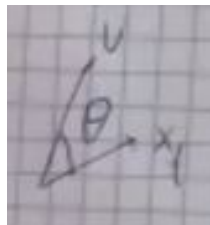


Figure 13.2: Tree building

where  $\Theta$  is the angle

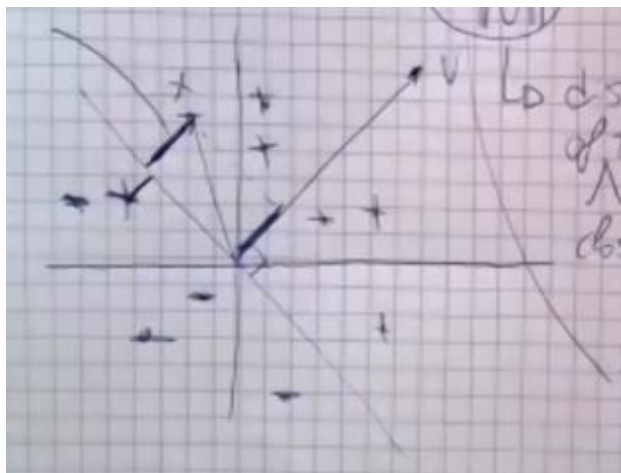


Figure 13.3: Tree building

where  $\frac{\gamma(U)}{\|U\|}$  is the distance separating hyperplane on closest training example .

S linearly separable and if i look at the sistem of this linear inequality:

$$\begin{cases} y_t w_T x_t > 0 \\ y_m w_T x_m > 0 \end{cases}$$

We can solve it in polytime using a linear solver. So any package of linear programming, and will be solved in linear time.

This is called **feasibility problem**. We want a point  $y$  that satisfy all my linear inequalities.

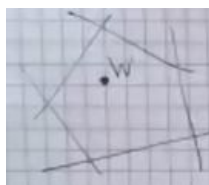


Figure 13.4: Feasibility problem

**When  $Opt(S) = 0$  is we can implememtn  $ERM$  efficiently using LP (Linear programming).**

They may overfitting since a lot of bias. When this condition of Opt is no satisfy we cannot do it efficiently. LP algorithm can be complicated so we figure out another family of algorithm.

## 13.2 The Perception Algorithm

This came from late '50s and was designed for psicology but have a general utility in othe fields.

Perception Algorithm

Input : training set  $S = \{(x_t, y_t) \dots (x_m, y_m)\}$       $x_t \in \mathbb{R}^d$       $y_t \in \{-1, +1\}$

Init:  $w = (0, \dots, 0)$

Repcat

read next  $(x_t, y_t)$

If  $y_t w^T x_t \leq 0$  then  $w \leftarrow w + y_t x_t$

Until margin greater than 0  $\gamma(w) > 0$  // w separates  $S$

Output  $w$

We know that  $\gamma(w) = \min_t y_t w^T x_t \leq 0$  The question is, will it terminate if  $S$  is linearly separable?

If  $y_t w^T x_t \leq 0$ , then  $w \leftarrow w + y_t x_t$

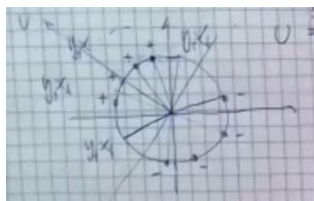


Figure 13.5:

For simplicity our  $x$  are in this circle. Some are on the circumference on top left with + sign and some in bottom right with - sign.

All minus flipped to the other side and the we can deal the +.

$U$  is a separating hyperplane, how can i find it?

Maybe i can do something like the average:

$$U = \frac{1}{m} \sum_{t=1}^m y_t x_t \quad ?$$

But actually don't take the average of all of them. So do not take average of



all, instead take the one that satisfy  $y_t w^T x_t \leq 0$  condition.  
 $y_t w^T x_t \leq 0$  is a violated constraint and we want it  $> 0$ .  
 Does  $w \leftarrow w + y_t x_t$  fix it?

$$y_t(w + y_t \cdot x_t)^T x_t = y_t w^T x_t + \|x_t\|^2$$

We are trying to see what happen before and after the updates of  $w$ .  
 Since  $\|x_t\| > 0$  so is positive, the update increase margins, thus going to-  
 wards fixing violated constraints.

### 13.2.1 Perception convergence Theorem

dated early 60s On a linearly separable  $S$ , perceptron will converge after at  
 most  $M$  updates (when they touch in the figure) where:

$$M \leq \left( \min_{U: \gamma(U)=1} \|U\|^2 \right) \left( \max_{t=1,..,m} \|x_t\|^2 \right)$$

Algorithm is not able to do that. Algorithm keeps looking till he get a vio-  
 lating constraint and then stops. This is bounded by the number of loops.

We said that  $\gamma(U) = \min_t y_t U^T x_t > 0$  when  $U$  is separator.

$$\forall t \quad y_t U^T x_t \geq \gamma(U) \quad \Leftrightarrow \quad \forall t \quad y_t \left( \frac{U}{\gamma(U)} \right)^T x_t \geq 1$$

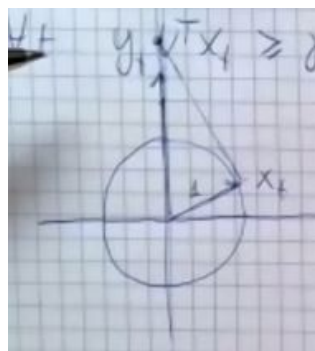


Figure 13.6:

If i rescale  $U$  i can make the margin bigger (in particular  $> 1$ )

The shortest  $\min \|U\|$  s.t.  $y_t U^T x_t \geq 1 \quad \forall t$

**Proof:**

$W_m$  is local variable after  $M$  updates, I have zero vector  $W_0 = (0, \dots, 0)$   
 $t_M$  is the index of training example that causes the  $M$ -th update.

We want to upper bound  $M$  (deriving upper and lower bound on a certain quantity  $\|W\| \|U\|$ )  
 where  $U$  is any s.t.  $y_t U^T x_t \geq 1 \quad \forall t$

$$\begin{aligned} \|W_M\|^2 &= \|W_{M-1} + y_{t_M} x_{t_M}\|^2 = \|W_{M-1}\|^2 + \|y_{t_M} x_{t_M}\|^2 + 2 \cdot y_{t_M} W_{M-1}^T x_{t_M} = \\ &= \|W_{M-1}\|^2 + \|x_{t_M}\|^2 + 2 \cdot y_{t_M} W_{M-1}^T x_{t_M} \leq \end{aligned}$$

where  $y_{t_M} W_{M-1}^T x_{t_M} \leq 0$

$$\leq \|w_{M-1}\|^2 + \|x_{t_M}\|^2$$

$$\|W_M\|^2 \leq \|W_0\|^2 + \sum_{i=1}^M \|x_{t_i}\|^2 \leq M \left( \max_t \|x_t\|^2 \right)$$

.....

.....

... MANCA ?????

$$\|W_M\| \|U\| \leq \|U\| \sqrt{M} \left( \max_t \|x_t\| \right)$$

since  $\cos \Theta \in [-1, 1]$

$$\|W_M\| \|U\| \geq \|W_M\| \|U\| \cos \Theta = W_M^T U = (W_{M-1} + y_{t_M} x_{t_M})^T U =$$

where last passage is the **Inner product**

$$W_{M-1}^T U + y_{t_M} U^T x_{t_M} \geq W_{M-1}^T U + 1 \geq W_0^T U + M = M$$

where  $y_{t_M} U^T x_{t_M}$  is  $\geq 1$

$$M \leq \|W_M\| \|U\| \leq \|U\| \sqrt{M} \left( \max_t \|x_T\| \right)$$

$$M \leq (\|U\|^2) \left( \max_t \|x_t\|^2 \right) \quad \forall U : \min_t y_t U^T x_t \geq 1$$

$$M = \left( \min_{U: \gamma(U)=1} \|U\|^2 \right) \left( \max_t \|x_t\|^2 \right)$$

Some number depends on  $S$

$M$  can be exponential in  $md$  when the ball of positive and negative are very closer and the length of  $U$  is super long and exponential in  $D$ .

If dataset barely separable then perceptron will make a number of mistakes that is exponential in the parameter of the problem.  $U$  is a linear separator and has exponential length

# Lecture 14 - 28-04-2020

## 14.1 Linear Regression

Yesterday we look at the problem of empirical risk minimisation for a linear classifier. 0-1 loss is not good: discontinuous jumping from 0 to 1 and it's difficult to optimise. Maybe with linear regression we are luckier.

Our data points are the form  $(x, y)$   $x \in \mathbb{R}^d$  regression,  $(\hat{y} - y)^2$  square loss. We are able to pick a much nicer function and we can optimise it in an easier way.

### 14.1.1 The problem of linear regression

Instead of picking -1 or 1 we just leave it as it is.

$$h(x) = w^T x \quad w \in \mathbb{R}^d \quad x = (x_1, \dots, x_d, 1)$$

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d} \frac{1}{m} \sum_{t=1}^m (w^T x_t - y_t)^2 \quad \text{ERM for } (x_1, y_1) \dots (x_m, y_m)$$

How to compute the minimum?

We use the vector  $v$  of linear prediction

$$v = (w^T x_1, \dots, w^T x_m)$$

and a vector of real valued labels

$$y = (y_1, \dots, y_m) \text{ where } v, y \in \mathbb{R}^m$$

$$\sum_{t=1}^m (w^T x_t - y_t)^2 = \|v - y\|^2$$

$S$  is a matrix.

$$S^T = [x_1, \dots, x_m] \quad d \times m \quad v = Sw = \begin{bmatrix} x_1^T \\ \dots \\ x_m^T \end{bmatrix} \begin{bmatrix} w \end{bmatrix}$$

So:

$$\|v - y\|^2 = \|Sw - y\|^2$$

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d} \|Sw - y\|^2 \quad \text{where } Sw \text{ is the design matrix}$$

$$F(w) = \|Sw - y\|^2 \quad \text{is convex}$$

$$\nabla F(w) = 2s^T (sw - y) = 0 \quad s^T sw = s^T y$$

where  $s^T$  is  $d \times m$  and  $s$  is  $m \times d$  and  $d \neq m$

If  $s^T s$  invertible (non singular)  $\hat{w} = (s^T s)^{-1} s^T y$

And this is called **Least square solutions (OLS)**

We can check  $s^T s$  is non-singular if  $x_1, \dots, x_m$  span  $\mathbb{R}^d$

$s^T \cdot s$  may not be always invertible. Also Linear regression is high bias solution. ERM may underfit since linear predictor introduce big bias.

$\hat{w} = (s^T \cdot s)^{-1} \cdot s^T \cdot y$  is very instable: can change a lot when the the dataset is perturbed.

This fact is called **instability** : variance error

It is a good model to see what happens and then try more sofisticated model.

Whenever  $\hat{w}$  is invertible we have to prove the instability. But there is a easy fix!

## 14.1.2 Ridge regression

We want to stabilised our solution. If  $s^T \cdot s$  non-singular is a problem.

We are gonna change and say something like this:

$$\hat{w} = \arg \min_w \|s \cdot w - y\|^2 \quad \rightsquigarrow \hat{w}_\alpha = \arg \min_w (\|s w - y\|^2 + \alpha \cdot \|w\|^2)$$

where  $\alpha$  is the **regularisation term**.

$\hat{w}_\alpha \rightarrow \hat{w}$  for  $\alpha \rightarrow 0$

$\hat{w}_\alpha \rightarrow (0, \dots, 0)$  for  $\alpha \rightarrow \infty$

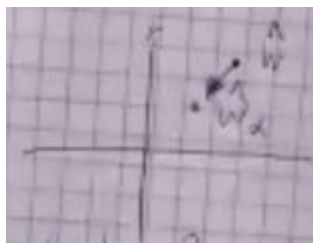


Figure 14.1:

$\hat{w}_\alpha$  has more bias than  $\hat{w}$ , but also less variance

$$\begin{aligned} \nabla (\|s w - y\|^2 + \alpha \|w\|^2) &= 2 (s^T s w - s^T y) + 2 \alpha w = 0 \\ (s^T s + \alpha I) w &= s^T y \\ (d \times m) (m \times d) (d \times d) (d \times m) &\quad (d \times m) (m \times 1) \end{aligned}$$

where  $I$  is the identity

$$\hat{w}_\alpha = (s^T s + \alpha I)^{-1} s^T y$$

where  $y_1, \dots, y_\alpha$  are eigen-values of  $s^T s$

$y_1, \dots, y_\alpha + \alpha > 0$  eigenvalues of  $s^T s + \alpha I$

In this way we make it positive and semidefinite.

We can always compute the inverse and it is a more stable solution and stable means **do not overfit**.

## 14.2 Perceptron

Now we want to talk about algorithms.

Data here are processed in a sequential fashion one by one.

Each datapoint is processed in constant time  $\Theta(d)$

(check  $y_t w^T \leq 0$  and in case  $w \leftarrow w + y_t x_t$ ) and the linear model can be stored in  $\Theta(d)$  space.

Sequential processing scales well with the number of datapoints.

But also is good at dealing with scenarios where new data are generated at all times.

Several scenarios like:

- Sensor data
- Financial data
- Logs of user

So sequential learning is good when we have a lot of data and a scenario in which data comes in fits like sensor.

We call it **Online learning**

## 14.2.1 Online Learning

It is a learning protocol and we can think of it like Batch learning. We have a class  $H$  of predictors and a loss function  $\ell$  and we have an algorithm that outputs an initial default predictor  $h_1 \in H$ .

For  $t = 1, 2, \dots$

- 1) Next example  $(x_t, y_t)$  is observed
- 2) The loss  $\ell(h_t(x_t), y_t)$  is observed  $(y_t w^T x_t \leq 0)$
- 3) The algorithm updates  $h_t$  generating  $h_{t+1}$   $(w \leftarrow w + y_t x_t)$

The algorithm generates a sequence  $h_1, h_2, \dots$  of models

It could be that  $h_{t+1} = h_t$  occasionally

The update  $h_t \rightarrow h_{t+1}$  is **local** (it only uses  $h_t$  and  $(x_t, y_t)$ )

This is a batch example in which take the training set and generate a new example.

$$(x_1, y_1) \rightarrow A \rightarrow h_2$$

$$(x_1, y_1)(x_2, y_2) \rightarrow A \rightarrow h_3$$

But if I have a non-learning algorithm I can look at the updates:

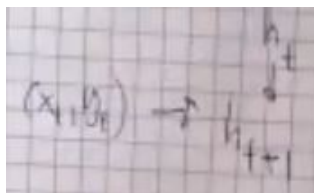


Figure 14.2:

This is a most efficient way and can be done in a constant time. The batch learning usually has a single predictor while the online learning uses a sequence of predictors.

How do I evaluate an online learning algorithm  $A$ ? I cannot use a single model, instead we use a method called **Sequential Risk**.

Suppose that I have  $h_1, h_2, \dots$  on some data sequence.

$$\frac{1}{m} \sum_{t=1}^T \ell(h_t(x), y_t) \quad \text{as a function of } T$$

The loss on the next incoming example.

I would like something like this:

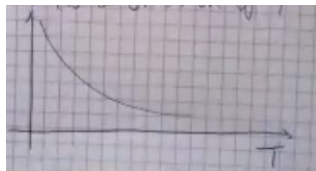


Figure 14.3:

We need to fix the sequence of data: I absorb the example into the loss of the predictor.

$$\ell(h_t(x), y_t) \longrightarrow \ell_t(h_t)$$

I can write the sequential risk of the algorithm:

$$\frac{1}{m} \sum_{t=1}^T \ell_t(h_t) - \min_{h \in H} \frac{1}{m} \sum_{t=1}^T \ell_t(h)$$

So the sequential risk of the algorithm - the sequential risk of best predictor in  $H$  (up to  $T$ ).

**This is a sequential similar of variance error.**  $\longrightarrow$  is called **Regret**.

$$h_T^* = \arg \min_{h \in H} \frac{1}{T} \sum_t \ell_t(h) \quad \frac{1}{T} \sum_t \ell_t(h_t) - \frac{1}{T} \sum_t \ell_t(h_T^*)$$



## 14.2.2 Online Gradient Descent (OGD)

It is an example of learning algorithm.

In optimisation we have one dimension and we want to minimise the function  $f$ . We can compute the gradient in every point.

We start from a point and get the derivative: as I get the derivative I can see if it is decreasing or increasing.

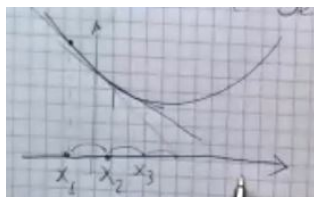


Figure 14.4:

$$\begin{aligned} f & \text{ convex} & \min_x f(x) & \quad f: \mathbb{R}^d \rightarrow \mathbb{R} \\ x_{t+1} & = x_t + \eta \nabla f(x_t) & \eta > 0 \\ w_{t+1} & = w_t + \eta \nabla \ell_t(w_t) \end{aligned}$$

where  $\eta$  is the learning rate.

$$h(x) = w^T x \quad \ell_t(w) = \ell(w^T x_t, y_t) \quad \text{for instance } \ell(w^T x_t, y_t) = (w^T x_t - y_t)^2$$

Assumption  $\ell_t$  is convex (to do optimisation easily) and differentiable (to compute the gradient)

# Lecture 15 - 04-05-2020

## 15.1 Regret analysis of OGD

We introduce the **Regret**.

$$\frac{1}{m} \sum_{t=1}^T \ell_t(w_t) - \frac{1}{T} \sum_{t=1}^T \ell_t(u_t^*)$$
$$(x_1, y_1) \dots (x_t, y_t) \quad \ell_t(w) = (w^T x_t - y_t)^2$$

we build a loss function for example with the square loss.

The important thing is that  $\ell_1, \ell_2, \dots$  is a sequence of **convex losses**.

In general we define the regret in this way:

$$R_T(u) = \sum_{t=1}^T \ell_t(w_t) - \sum_{t=1}^T \ell_t(u_t)$$

The Gradient descent is one of the simplest algorithm for minimising a convex function. We recall the iteration did by the algorithm:

$$w_{t+1} \leftarrow w_t - \eta_t \nabla f(w_t) \quad \eta_t > 0 \text{ **learning rate** } \quad f \text{ convex}$$

$f : \mathbb{R}^d \rightarrow \mathbb{R}$  that's why use the gradiand instead of the derivative

Learning rate can depend on time and we approach the region of the function f where the region is 0. We keep on moving in the X axes in the direction where the function is decreasing.

### 15.1.1 Projected OGD

2 parameters:  $\eta > 0$  and  $U > 0$

Initialisation:  $w_1 = (0, \dots, 0)$

For  $t = 1, 2, \dots$

1) **Gradient step:**

$$w'_{t+1} = w_t - \frac{\eta}{\sqrt{t}} \nabla \ell_t(w_t) \quad (x_t, y_t) \rightsquigarrow \ell_t$$

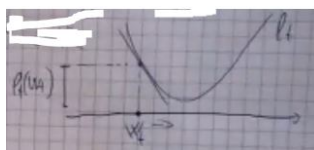


Figure 15.1:

2) **Projection step:**

$$w_{t+1} = \arg \min_{w: \|w\| \leq U} \|w - w'_{t+1}\|$$

**Projection of  $w'_{t+1}$  onto the ball of radius  $U$ .**

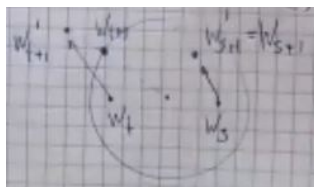


Figure 15.2:

Now we define the Regret:

$$U_T^* = \arg \min_{U \in \mathbb{R}^d, \|U\| \leq U} \frac{1}{T} \sum_{t=1}^T \ell_t(U)$$

We are interested in bounding the regret  $R_T(U_T^*)$

I will fix  $\ell_1, \dots, \ell_t$  let  $U = U_T^*$  for  $U$ .

Taylor's theorem for multivariate functions

Let's look a univariate first  $f : \mathbb{R} \rightarrow \mathbb{R}$  ( has to be twice differentiable)  
 $w, u \in \mathbb{R}$

$$f(u) = f(w) + f'(w)(u - w) + \frac{1}{2} f''(x)(u - w)^2$$

For the multivariate case:

$f : \mathbb{R}^d \rightarrow \mathbb{R}$  twice differentiable  $\forall u, w \in \mathbb{R}^d$

$$f(u) = f(w) + \nabla f(w)^T (u - w) + \frac{1}{2} (u - w)^T \nabla^2 f(\xi) (u - w)$$

where  $\xi$  is some point on the segment going  $u$  and  $w$ . We have the Hessian matrix of  $f$ :

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} |_{x = x_i}$$

If  $f$  is convex then,  $\nabla^2 f$  is positive and semidefinite.  
 $\forall x \in \mathbb{R}^d \quad \forall z \in \mathbb{R}^d \quad z^T \nabla^2 f(x) z \geq 0$



Figure 15.3:

Now we can apply this results to our problem: in particular I rearrange the factors

$$f(w) - f(u) \leq \nabla f(w)^T (w - u)$$

This is Ok for  $f$  convex and differentiable.

I know that:  $u - w^T \nabla^2 f(\xi) (u - w) \geq 0$  because  $f$  is convex.

$$\ell_t(w_t) - \ell_t(u) \leq \nabla \ell_t(w_t)^T (w_t - u) \quad \text{Linear Regret}$$

How do we proceed?

The first step of the algorithm is :  $w'_{t+1} = w_t - \eta_t \nabla \ell_t(w_t) \quad \eta_t = \frac{\eta}{\sqrt{t}}$

$$= -\frac{1}{\eta_t} (w'_{t+1} - w_t)^T (w_t - u) = \frac{1}{\eta_t} \left( \frac{1}{2} \|w_t - u\|^2 - \frac{1}{2} \|w'_{t+1} - u\|^2 + \frac{1}{2} \|w_{t+1} - w_t\|^2 \right) \leq$$

$$\leq \frac{1}{\eta_t} \left( \frac{1}{2} \|w_t - u\|^2 - \frac{1}{2} \|w_{t+1} - u\|^2 + \frac{1}{2} \|w'_{t+1} - w_t\|^2 \right)$$

$w'$  disappear and add minus sign. I am saying that  $\|w_{t+1} - u\| \leq \|w'_{t+1} - u\|$

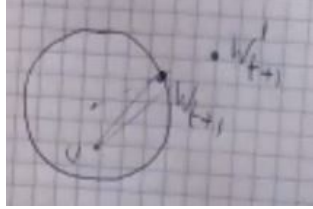


Figure 15.4:

So is telling us that  $w_{t+1}$  is closer to  $u$  than  $w'_{t+1}$   
This holds since the ball is convex.

Now we go back adding and subtracting  $\pm \frac{1}{2\eta_{t+1}} \|w_{t+1} - u\|^2$

$$= \frac{1}{2\eta_t} \|w_t - u\|^2 - \frac{1}{2\eta_{t+1}} \|w_{t+1} - u\|^2 - \frac{1}{2\eta_t} \|w_{t+1} - u\|^2 + \frac{1}{2\eta_{t+1}} \|w_{t+1} - u\|^2 + \frac{1}{2\eta_t} \|w_{t+1} - w_t\|^2$$

We group the 1,2 and 3,4 elements and sum them up.

$$R_T(U) = \sum_{t=1}^T (\ell_t(w_t) - \ell_t(u)) \leq$$

This is a **telescopic sum**:  $a_1 - a_2 + a_2 - a_3 + a_3 - a_4 + \dots + a_t - a_{t+1} + 1$  and everything in the middle cancel out and remains first and last terms.

$$\leq \frac{1}{2\eta_t} \|w_1 - u\|^2 - \frac{1}{\eta_{T+1}} \|w_{T+1} - u\|^2 + \frac{1}{2} \sum_{t=1}^T \|w_{t+1} - u\|^2 \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \frac{1}{2} \sum_{t=1}^T \frac{\|w'_{t+1} - w_t\|^2}{\eta_t}$$

where  $w_1 = 0$  and  $\|w_{t+1} - u\|^2 \leq 4U^2$  and  $\|w'_{t+1} - w_t\|^2 = \eta_t^2 \|\nabla \ell_t(w_t)\|^2$   
We know that  $\eta_t = \frac{\eta}{\sqrt{t}}$  so  $\eta_1 = \frac{\eta}{\sqrt{1}} = \eta$

$$R_T(U) \leq \frac{1}{2\eta} U^2 - \frac{1}{2\eta_{T+1}} \|w_{T+1} - U\|^2 + 2U^2 \sum_{t=1}^{T-1} \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t+1}} \right) + \frac{\|w_{T+1} - U\|^2}{2\eta_{T+1}} - \frac{\|w_T - U\|^2}{\eta_T} + \frac{1}{2} \sum_{t=1}^T \eta_t \|\nabla \ell_t(w_t)\|^2$$

where **red values** cancel out.

I assume that square loss is bounded by some number  $G^2$ :  $\|\nabla \ell_t(w_t)\|^2 \leq G^2$

Also, it's a telescopic sum again and all middle terms cancel out.

$$\max_t \|\nabla_t(w_t)\|^2 \leq G$$

$$R_T(U) \leq \frac{1}{2\eta} U^2 + 2U^2 \left( \frac{1}{\eta_T} - \frac{1}{\eta_1} \right) + \frac{G^2}{2} \eta \sum_{t=1}^T \frac{1}{\sqrt{t}} \quad \eta_t = \frac{1}{\sqrt{t}}$$

where **red values** cancel out.

Now how much is this sum  $\sum_{t=1}^T \frac{1}{\sqrt{t}}$ ?

It is bounded by the integral  $\leq \int_1^T \frac{dx}{\sqrt{x}} \leq 2\sqrt{T}$

$$R_T(U) \leq \frac{2U^2\sqrt{T}}{\eta} + \eta G^2\sqrt{T} = \left( \frac{2U^2}{\eta} + \eta G^2 \right) \sqrt{T}$$

$$\eta = \frac{U}{G\sqrt{2}}$$

So finally:

$$\frac{1}{T} \sum_{t=1}^T \ell_t(w_t) \leq \min_{\|u\| \leq U} \frac{1}{T} \sum_{t=1}^T \ell_t(u) + U G \sqrt{\frac{8}{T}}$$

$$R_T(U) = \frac{1}{T} \sum_{t=1}^T (\ell_t(w_t) - \ell_t(u)) \quad \forall u : \|u\| \leq U : R_T(U) = O\left(\frac{1}{\sqrt{T}}\right)$$

Basically my regret is gonna go to 0.

For *ERMinH* where  $|H| < \infty$ , variance error vanishes at rate  $\frac{1}{\sqrt{m}}$

The bound  $U G^2 \sqrt{\frac{8}{T}}$  on regret holds for any sequence  $\ell_1, \ell_2, \dots$  of convex and affordable losses, If  $\ell_t(w) = \ell(w^T x_t, y_t)$  then the bound holds for any sequence of data points  $(x_1, y_1), (x_2, y_2) \dots$ .

This is not a statistical assumption but mathematical so stronger.

# Lecture 16 - 05-05-2020

## 16.1 Analysis of Perceptron in the non-separable case using OGD framework.

We are finishing up the part of online learning and gradient descent.  
Concrete example for the parameter  $G$ :

$$G = \max_t \|\nabla \ell_t(w_t)\|^2 \quad \ell_t(w) = (w^T x_t - y_t)^2 \quad \|x_t\| \leq X, |y_t| \leq UX$$

$$\|w\| \leq U, |w^T x_t| \leq \|w\| \|x_t\|$$

where  $\|w\| \rightsquigarrow U$  and  $\|x_t\| \rightsquigarrow X$  (so are bounded by  $U$  and  $X$ )

Now we want to find the gradient.

$$\|\nabla \ell_t(w)\| \leq 2 |w^T x_t - y_t| \|x_t\| \leq 4UX \|x_t\| \leq 4UX^2$$

where  $w^T x_t$  bounded  $UX$  and  $y_t$  bounded by  $UX$

$$R_T(u) = UG\sqrt{8T} \leq 8(UX)^2\sqrt{2T}$$

How about OGD for classification?

The problem is that zero-one loss is not convex (also non-continuous).

$$I\{y_t w^T x_t \leq 0\}$$

This is zero-one loss for linear classification.

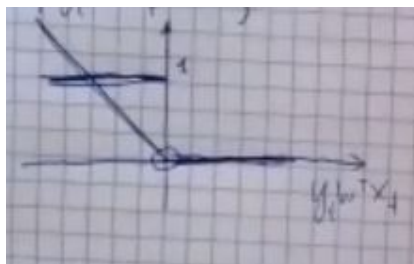


Figure 16.1:

$$w \leftarrow w - \eta \nabla \ell_t(w) \quad \text{OGD}$$

$$w \leftarrow w + y_t x_t I\{y_t w^T x_t \leq 0\} \quad \text{Perceptron}$$

We want to make this equal with loss that is convex and also tell us bound with zero one loss. So there is a bunch or problem.

So we want to make this equal but how?

$$\ell_t(w) = [-y_t w^T x_t]_+ \quad [z]_+ = \max\{0, z\}$$

If we take the gradient of this with respect to  $w$ :

$$\nabla \ell_t(w) = -y_t x_t I\{y_t w^T x_t \leq 0\}$$

Now - this gradient is exactly this  $\ell_t(w) = [-y_t w^T x_t]_+$

The problem is not comparable with the number of mistakes so I am not going to have the number of mistakes.

How do I do it?

What if I just shift to the right?

Now this loss is an upper bound of the zero-one loss. And this is called **Hinge loss** (where hinge take the door attached to the frame of the wall)

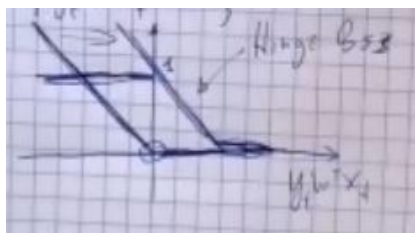


Figure 16.2: Hinge loss

**Hinge loss:**  $h_t(w) = [1 - y_t w^T x_t]_+ \geq I\{y_t w^T x_t \leq 0\}$

$$\nabla h_t(w) = -y_t x_t I\{y_t w^T x_t \leq 1\}$$

The problem is that it becomes 0 later on than the original one.

$$w \leftarrow w - \eta \nabla h_t(w) I\{y_t w^T x_t \leq 0\}$$

$$y_t w^T x_t \leq 0 \Rightarrow y_t w^T x_t \leq 1$$

We now apply OGD analysis to  $h_t$  considering only the steps  $T$  where  $I\{y_t w_t^T x_t \leq 0\}$  and we do not perform projection.

$$\sum_{t=1}^T (h_t(w_t) - h_t(u)) I\{y_t w_t^T x_t \leq 0\} \leq$$



$$\leq \frac{1}{2\eta} \|U\|^2 + \frac{1}{2} \sum_{t=1}^T \|w_{t+1} - u\| \left( \frac{1}{\eta} - \frac{1}{\eta} \right) I\{y_t w_t^T x_t \leq 0\} + \frac{\eta G^2}{2} \sum_{t=1}^T I\{y_t w_t^T x_t \leq 0\}$$

where second factor cancel out

$$-\frac{1}{2\eta} \|w_{T+1} - u\|^2 \quad \|\nabla h_t(w)\| = |y_t| \|x_t\| \leq X \quad G = X = \max_t \|x_t\|$$

where  $y_t$  in  $\{-1, 1\}$

$$y_t w_t^T x_t \leq 0 \Rightarrow h_t(w_t) \geq 1$$

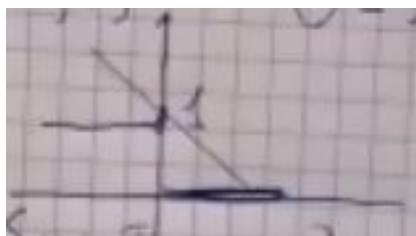


Figure 16.3:

$$\begin{aligned} \sum_{t=1}^T I\{y_t w_t^T x_t \leq 0\} &\leq \sum_{t=1}^T h_t(w_t) I\{y_t w_t^T x_t \leq 0\} \leq \\ &\sum_{t=1}^T h_t(u) I\{y_t w_t^T x_t \leq 0\} + \frac{1}{2\eta} \|u\|^2 + \frac{\eta}{2} x^2 \sum_{t=1}^T I\{y_t w_t^T x_t \leq 0\} \end{aligned}$$

where  $I..$  cancel out to have a "nicer" upper bound.

$$M_T = \sum_{t=1}^T I\{y_t w_t^T x_t \leq 0\}$$

$$M_T \leq \sum_{t=1}^T h_t(u) + \frac{1}{2\eta} \|u\|^2 + \frac{\eta}{2} x^2 M_T$$

This is not a regret anymore! Here  $M_T$  is the number of mistakes and I compare it with the hinge loss ( $h_t(u)$ ).

I CAN'T USE THIS  $\eta = \frac{\|u\|}{x\sqrt{M_T}}$  but we can replace it in  $M_T$ .

$$M_T \leq \sum_{t=1}^T h_t(u) + \|u\| X \sqrt{M_T}$$

$$M_T \leq \sum_{t=1}^T h_t(u) + (\|u\|x)^2 + \|u\|x \sqrt{\sum_t h_t(u)}$$

$$w \leftarrow w + \eta y_t x_t I\{y_t w_t^T x_t \leq 0\} \quad w = (0, \dots, 0)$$

If I choose  $\eta > 0$ ?

$$w_t = \eta \sum_{s=1}^{t-1} y_s x_s I\{y_s w_s^T x_s \leq 0\} \quad \forall \eta > 0$$

Also holds because it's true  $\forall \eta > 0$

$$M_T = \sum_{t=1}^T I\{y_t w_t^T x_t \leq 0\} \quad \text{invariant with respect to } \eta > 0$$

It does not matter which  $\eta$  we choose. The number of mistakes is going to be the same. This means that the state of the algorithm (which depends on mistakes) is gonna be the same.

I can run perceptron with  $\eta = 1$  and pretend (in the analysis) it was run with  $\eta = \frac{\|U\|}{X\sqrt{M_T}}$

We go back to the bound of  $M_T$ . We are actually free to choose any number of  $U$ .

If  $(x_1, y_1), (x_2, y_2)$  is linearly separable then:

$$\exists U \text{ s.t. } y_t U^T x_t \geq 1 \Rightarrow h_t(u) = 0 \quad \forall t$$

$$M_T \leq (\|U\| X)^2 \quad \text{the **perceptron convergence theorem** .}$$

$$M_T \leq \min_{u \in \mathbb{R}^d} \left( \sum_{t=1}^T h_t(u) + (\|U\| X)^2 + \|U\| X \sqrt{\sum_t h_t(u)} \right)$$

This are called **Oracle bounds**, the perceptron knows which is the best  $U$ .

### 16.1.1 Strongly convex loss functions

We use this to analyse all class of algorithms that regularise the ERM which is the support vector machine. We want to explain what happen using Support vector Machine. For neural networks we cannot do this since NN are not convex and there is not way to "convexify". Convexifying we lose the power of NN.

We said that  $\ell_t$  have to be convex. But i have a lot of types of convexity.

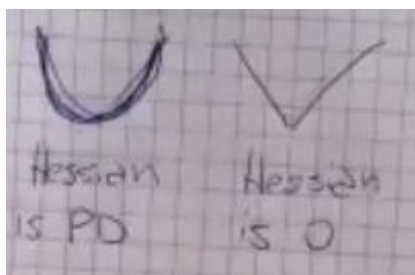


Figure 16.4: Example of more type of convex function

This two for example are both convex. In the left this always has a positive curvature, while the right one we have a 0 curvature since is two straight line and not differentiable. In other word, Hessian on the left positive and definite. On the right Hessian is 0.

We are looking for **strongly convex losses**.

$\ell$  differentiable is  $\sigma$ -strongly convex if:

$$\forall u, w \quad \ell(w) - \ell(u) \leq \nabla \ell(w)^T (w - u) - \frac{\sigma}{2} \|w - u\|^2 \quad \sigma > 0$$

$\sigma$ -SC is equivalent to the Hessian having all strictly positive eigenvalues.

Example, check if strictly convex:

$$\ell(w) = \frac{1}{2} \|w\|^2 \quad \frac{1}{2} \|w\|^2 - \frac{1}{2} \|u\|^2 \stackrel{?}{\leq} w^T (w - u) - \sigma \frac{\|w - u\|^2}{2}$$

where  $\sigma = 1$

$$\frac{1}{2} \|w\|^2 - \frac{1}{2} \|u\|^2 \stackrel{?}{\leq} \|w\|^2 - w^T u - \frac{\|w - u\|^2}{2}$$

where  $\frac{1}{2} \|w\|^2$  cancel out

$$0 \leq \frac{1}{2} \|w\|^2 + \frac{1}{2} \|u\|^2 - \frac{\|w - u\|^2}{2} - w^T u$$

I put  $0 = \dots$

$$0 = \frac{1}{2}\|w\|^2 + \frac{1}{2}\|u\|^2 - \frac{\|w - u\|^2}{2} - w^T u$$

So this function is *1-strongly convex*

Next lecture we are going to show that we can run OGD with strongly convex functions. We are going to get a better bound. Our regret is gonna vanish much faster than the case of simple convexity.

You can prove that if Hessian is  $\mu$ , your regret is vanishing with a rate of  $\frac{UG}{\sqrt{T}}$ . We will show with strong convexity the OGD will converge much faster with a rate of  $\frac{\ln T}{T}$ .

This is what happens in optimisation, we prefer strictly convex function.

# Lecture 17 - 11-05-2020

## 17.1 Strongly convex loss functions

We will saw with OGD but we will see Support Vector Machine(SVM). Very popular learning model.

We will see SVM next to see the part of linear predictor and also speak about Kernel function used with linear predictor to obtain non-linear classifier from a linear classifier.

$\ell$  is  $\sigma$ -SC if  $\forall u, w$ :

$$\ell(w) - \ell(u) \leq \nabla \ell(w)^T (w - u) - \frac{\sigma}{2} \|w - u\|^2$$

### 17.1.1 OGD for Strongly Convex losses

Init:  $w_1 = (0, \dots, 0)$

For  $t = 1, 2, \dots$

$$w_{t+1} = w_t - \frac{1}{\sigma t} \nabla \ell_t(w_t) \quad \eta_t = \frac{1}{\eta t}$$

(no projection steps)

$$\begin{aligned} \ell_t(w_t) - \ell_t(u) &\leq \nabla \ell_t(w_t)^T (w_t - u) - \frac{\sigma}{2} \|w_t - u\|^2 = \\ &= -\frac{1}{\eta_t} (w_{t+1} - w_t)^T (w_t - u) - \frac{\sigma}{2} \|w_t - u\|^2 = \\ &= \frac{1}{\eta_t} \left( \frac{1}{2} \|w_t - u\|^2 - \frac{1}{2} \|w_{t+1} - u\|^2 + \frac{1}{2} \|w_{t+1} - w_t\|^2 \right) - \frac{\sigma}{2} \|w_t - u\|^2 \\ R_T(u) &\leq \frac{1}{2\eta_1} \|w_1 - u\|^2 - \frac{1}{2\eta_{T+1}} \|w_{T+1} - u\|^2 - \frac{\sigma}{2} \|w_1 - u\|^2 + \\ &+ \frac{1}{2} \sum_{t=1}^{T-1} \|w_{t+1} - u\|^2 \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \sigma \right) + \frac{1}{2} \|w_{T+1} - u\|^2 \left( \frac{1}{\eta_{T+1}} - \frac{1}{\eta_T} \right) + \frac{G^2}{2} \sum_{t=1}^T \eta_t \end{aligned}$$

where red terms cancel out, **blue** (sum) instead is 0 since  $\sigma(t+1) - \sigma t - \sigma$

$$G = \max_t \|\nabla \ell_t(w_t)\|$$

$$R_T(U) \leq \frac{1}{2} (\sigma - \sigma) \|w_1 - u\|^2 + \frac{G^2}{2} \sum_{t=1}^T \frac{1}{\sigma t} =$$

$$R_T(U) \leq \frac{G^2}{2} \sum_{t=1}^T \frac{1}{\sigma t}$$

We know that  $\sum_{t=1}^T \frac{1}{t} \leq \ln(T+1)$  so:

$$R_T(U) \leq \frac{G^2}{2\sigma} \ln(T+1)$$

$\frac{R_T(U)}{T}$  **vanishes at rate**  $\frac{\ln T}{T} \ll \frac{1}{\sqrt{T}}$  **provided**  $\max_t \|\nabla \ell_t(w_t)\|$  **remains bounded**

We assume it in special case.

Where are these SC losses?

Minimising strongly convex version of standard convex losses helps a lot.

We will see how Regularisation imply Stability. Before studying SVM and stability we going to do something before.

## 17.1.2 Relate sequential risk and statistical risk

It is important: I have this algorithm that control sequential risk and regret but I am also curious to use this algorithms.

We assume:

Data  $(x_t, y)$  drawm i.i.d. from fixed unknown  $D$ .

Convex loss function  $\ell$ .

For example compare square loss and hinge loss(convex upper bound on 0-1 loss:

$$\ell(\hat{y}, y) = (\hat{y} - y)^2 \quad \ell(\hat{y}, y) = [1 - \hat{y}y]_+$$

We will focus on linear predictors  $h(x) = f(w^T x)$  (easily to analyse with OGD framework).

Risk  $\ell_D(w) = \mathbb{E}[\ell(w^T X, Y)]$

where  $\hat{y} = w^T X$

Assume we have a training set  $S$  of example  $(X_1, Y_1) \dots (X_m, Y_m)$  (in maiousc since are random sequence of data point from a distribution)

$$\text{Convex } \ell_t(w) = \ell(w^T X_t, Y_t) \quad t = 1, \dots, m$$

Became a sequence of convex losses.

I run OGD on  $\ell_1, \ell_2, \dots, \ell_m$  and get  $w_1, \dots, w_m$   $\|w_t\| \leq U$

OGD projects onto:

$$\{U \in \mathbb{R}^d : \|u\| \leq U\} \quad U^* = \arg \min_{u: \|u\| \leq U} \ell_D(u)$$

where  $U^*$  is the best linear predictor in class.

So I take a bunch of predictors but I need one, so I take the average of those (since the expected value is convex):

$$\bar{w} = \frac{1}{m} \sum_{t=1}^m w_t$$

I want to study the variance error:

$$\ell_D(\bar{w}) - \ell_D(u^*) ?$$

I am using Online Learning.

Using Jensen inequality:

$$\ell_D(\bar{w}) = \mathbb{E} [\ell(\bar{w}^T X, Y)] \leq \mathbb{E} \left[ \frac{1}{m} \sum_{t=1}^m \ell(w_t^T X, Y) \right] = \frac{1}{m} \sum_t \mathbb{E} [\ell(w_t^T X, Y)]$$

where  $\mathbb{E} [\ell(w_t^T X, Y)]$  is equals to  $\ell_D(w_t)$

$$\ell_D(\bar{w}) \leq \frac{1}{m} \sum_{t=1}^m \ell_D(w_t) \quad \text{for any given training set } (x_1, y_1) \dots (x_m, y_m)$$

I want to look at the difference:

$$\ell_D(w_t) - \ell(w_t^T X_t, Y_t)$$

$$\ell_D = \mathbb{E} [\ell(w_t^T X, Y)]$$

Now I fix  $t - 1$  example in the training set  $(X_1, Y_1) \dots (X_{t-1}, Y_{t-1})$

$w_t$  is **determined** by  $(X_1, Y_1), \dots (X_{t-1}, Y_{t-1})$

$(X_t, Y_t)$  is distributed like any  $(X, Y) \sim D$

$$\mathbb{E}_{t-1} [\cdot] = \mathbb{E} [\cdot | (X_1, Y_1) \dots (X_{t-1}, Y_{t-1})] \quad z_t = \ell_D(w_t) - \ell(w_t^T X_t, Y_t)$$

$$\frac{1}{m} \sum_{t=1}^m \mathbb{E}_{t-1} [Z_t] = 0$$

I want to show the average of  $\ell_D(w_t)$  is equal to average of  $\ell(w_t^T X_t, Y)$

I want to prove:

$$\frac{1}{m} \sum_{t=1}^m \ell_D(w_t) \leq \frac{1}{m} \sum_{t=1}^m \ell(w_t^T X_t, Y_t) + \sqrt{\frac{1}{m} \ln \frac{1}{\delta}} \quad \text{with high probability w.r.t. } S$$

where (red part) is the sequential risk of OGD.

$$\frac{1}{m} \sum_{t=1}^m Z_t \leq \sqrt{\frac{1}{m} \ln \frac{1}{\delta}} \quad \text{with prob. at least } 1 - \delta$$

I know that  $\mathbb{E}_{t-1}[Z_t] = 0$

$$|Z_t| \in [0, M] \quad \Rightarrow \quad \frac{1}{m} \sum_{t=1}^m Z_t \leq M \sqrt{\frac{2}{m} \ln \frac{1}{\delta}} \quad w.p. 1 - \delta$$

Version of Chernoff-Hoffdiwg bounds for sums of dependent random variables.

$$\frac{1}{m} \sum_{t=1}^m \ell_D(w_t) \leq \frac{1}{m} \sum_{t=1}^m \ell_t(w_t) + M \sqrt{\frac{2}{m} \ln \frac{1}{\delta}} \quad w.p. 1 - \delta$$

This tells me that  $\ell_D(\bar{w})$  is controlled by the sequential risk of OGD +  $O\left(\frac{1}{\sqrt{m}}\right)$   
 Variance Error for  $(w^T x - y)^2 \quad \|x_t\| \leq X, \quad |y_t| \leq U X$

$$G = \max_t \|\nabla \ell_t(w_t)\| \leq 4 (U X)^2$$

$$\ell_D(\bar{w}) \leq \min_{u: \|u\| \leq U} \frac{1}{m} \sum_{t=1}^m \ell_D(u) + 8 (U X)^2 \sqrt{\frac{2}{m}} + 4 (U X)^2 \sqrt{\frac{2}{m} \ln \frac{1}{\delta}}$$

where red is **OGD analysis**

$$\ell_D(\bar{w}) \leq \min \frac{1}{m} \sum_{t=1}^m \ell_t(u) + 12 (U X)^2 \sqrt{\frac{2}{m} \ln \frac{1}{\delta}} \quad \text{with prob. } 1 - \delta$$

By C-H bounds:

$$\text{where } \min \frac{1}{m} \sum_{t=1}^m \ell_t(u) \leq \frac{1}{m} \sum_{t=1}^m \ell_t(u^*) \leq \ell_D(u^*) + 4 (U X)^2 \sqrt{\frac{1}{2m} \ln \frac{1}{\delta}}$$

where the sum is the test error of  $u^*$   
 At the end:

$$\ell_D(\bar{w}) \leq \ell_D(u^*) + 16 (U X)^2 \sqrt{\frac{1}{m} \ln \frac{1}{\delta}} \quad w.p. 1 - \delta$$

Even with m large, I can run it since i bounded in the small "ball".



# Lecture 18 - 12-05-2020

## 18.1 Kernel functions

We use a notion of **feature expansion**. They are different but somehow they reach something similar. In fact Linear classifier have high bias. Linear predictor use hyper plane as basic brick to build prediction.

### 18.1.1 Feature expansion

Given  $\phi : \mathbb{R}^d \rightarrow V$   $V$  is typically  $\mathbb{R}^N$   $N \gg d$

For example:

$$d = 2 \quad N = 6 \quad \phi = \mathbb{R}^2 \rightarrow \mathbb{R}^6$$

$$\phi(x_1, x_2) = (1, x_1^2, x_2^2, x_1, x_2, x_1x_2)$$

We have a homogenous hyper plane.

$$w \in \mathbb{R}^6 \quad \{z \in \mathbb{R}^6 : w^T z = 0\} \quad z = \phi(x) \quad x \in \mathbb{R}^2$$

$$\forall x \in \mathbb{R}^2 \quad w^T \phi(x) = w_1 + w_2 x_1^2 + w_3 x_2^2 + w_4 x_1 + w_5 x_2 + w_6 x_1 x_2 = 0$$

$$w^T \phi(x) = 0$$

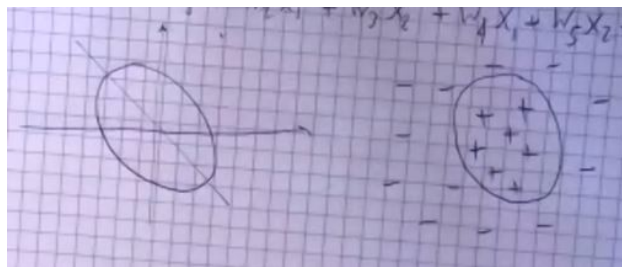


Figure 18.1:

$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^N \quad \prod_{s=1}^M x_{V_s} \quad v \in \{1, \dots, d\}^k \quad k = 0, \dots, n$$

$$h(x) = \text{sgn}(w^T \phi(x)) \quad w^T \phi(x) = \sum_{i=1}^N w_i \phi(x)_i$$

The problem of this feature expansion is the degree of the monomials!

$$N = \sum_{i=0}^n |\{1, \dots, d\}^k| = \sum_{k=0}^n d^k = \frac{d^{n+1} - 1}{d - 1} = \Theta(d^n)$$

So it's exponential! But this feature expansion can be implemented in a efficient way.

### 18.1.2 Kernels implements feature expansion (Efficiently)

$w^T \phi(x)$  Perceptron  $w \leftrightarrow w + y_t x_t I\{y_t w^T x_t \leq 0\}$  MANCA quadcosa

$$w = \sum_{s \in S} y_s x_s \rightsquigarrow \sum_{s \in S} y_s \phi(x_s)$$

where  $S$  is a subset of training set where updates occurred.

Every time i make a mistake i add some of this product of data points.

If I run this using example that are images according to some feature expansion map ( $\phi$ ), I will get the perceptron after the mapping.

$$w^T \phi(x) = \sum_{s \in S} y_s \phi(x)^T \phi(x_s)$$

It's a inner problem and can have exponentially degree of the component.

**Kernels help me compute this inner product  $\phi(x)^T \phi(x_s)$  quickly**

$$\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^6 \quad \phi(x_1, x_2) = (1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2)$$

$$\phi(x)^T \phi(z) = 1 + x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_2 z_2 + 2x_1 x_2 z_1 z_2 = (1 + x^T z)^2 = k(x, z)$$

$$w^T \phi(x) = \sum_{s \in S} y_s k(x, x_s)$$

$k(x, z)$  implements  $\phi(x)^T \phi(z)$   $\forall x, z$  and  $\phi$  defined as before

How to we generalise this?

$$k_n(x, x') = (1 + x^2 x')^n$$

This is called polynomial kernel.

I want to check now what is the  $\phi$  for  $K_n$ ?

I want to compute  $\phi$  s.t.  $\phi(x)^T \phi(x') = k_n(x, x') = (1 + x^T x')^n$

We can use Newtons binomial theorem:

$$(1 + x^T x')^n = \sum_{k=0}^n \binom{n}{k} (x^T x')^k$$

$$(x^T x')^k = \left( \sum_{i=1}^d x_i x'_i \right)^k = \sum_{v \in \{1, \dots, d\}^k} \left( \prod_{s=1}^k x_{V_s} x'_{V_s} \right)$$

$$\phi(x) = \left( \sqrt{\binom{n}{k}} \prod_{s=1}^k x_{V_s} \right) \quad k = 0, \dots, n \quad v \in \{1, \dots, d\}^k$$

When I am using polynomial kernel I am implicitly using the feature expansion ...

Can an algorithm work using kernel?

Perceptron works!

$S = 0$

For  $t = 1, 2, \dots$

1) Get  $(x_t, y_t)$

2)  $\hat{y}_t = \text{sgn} \left( \sum_{s \in S} y_s K(x, x_s) \right)$

3) If  $\hat{y}_t \neq y_t$   $S \leftarrow S \cup \{t\}$   $w \leftarrow w + y_t \phi(x_t)$

So I am representing  $y$  as a sum and not as a vector. In fact,  $w = \sum_{s \in S} y_s \phi(x_s)$

ù

## 18.2 Gaussian Kernel

$$\gamma > 0 \quad k_\gamma(x, x') = \exp \left( -\frac{1}{2\gamma} \|x - x'\|^2 \right)$$

$$e^{-\frac{1}{2\gamma} (x-x')^2}$$

I can controll the distribution changing the value of  $\gamma$

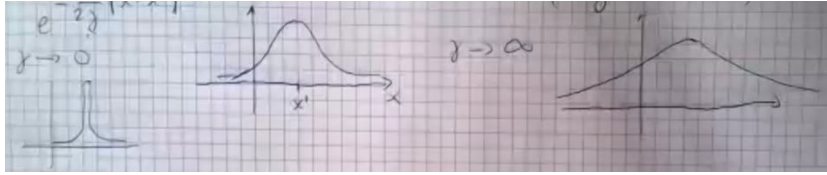


Figure 18.2:

$$\hat{y}_t = \text{sgn} \left( \sum_{s \in S} y_s K_\gamma(x, x_s) \right)$$

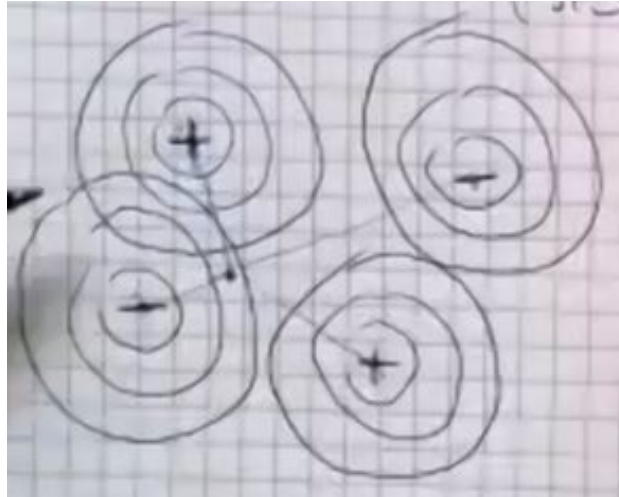


Figure 18.3:

Negative or positive gaussian component looking at the distance.

Now I want to compute:  $\phi_\gamma : \mathbb{R}^n \rightarrow V$

$$\exp \left( -\frac{1}{2\gamma} \|x - x'\|^2 \right) = \exp \left( -\frac{1}{2\gamma} (\|x\|^2 + \|x'\|^2) \right) \cdot \exp \left( \frac{1}{\gamma} x^T x' \right) =$$

where  $e = x + \frac{x^2}{12} \dots$

$$= \exp \left( -\frac{1}{2\gamma} \|x\|^2 \right) \cdot \exp \left( -\frac{1}{2\gamma} \|x'\|^2 \right) \cdot \sum_{n=0}^{\infty} \frac{1}{n!} \frac{(x^T x')^n}{\gamma^n}$$

Gaussian Kernel is a linear combination of infinitely many poly kernels.  
 The higher I go the small is  $n!$ . Gaussian kernel mapping into a space that is very large. So large that it has infinitely many dimension. Why? Because each polynomial kernel maps to infinitely dimensions.

$\phi_\gamma$  maps  $\mathbb{R}^d$  into a space of infinitely many dimensions.

$$\phi_\gamma : \mathbb{R}^d \rightarrow V \quad k_\gamma(x, x') = \phi_\gamma(x)^T \phi_\gamma(x')$$

It maps to infinitely many dimension, so it maps to a function!

$\phi_\gamma(x)$  is a function.

In general, when I learn a linear predictor using  $k_\gamma$

I learn  $\sum_s \alpha_s k(x_s, \cdot) = f$   
 $w^T \phi(x)$

$$H_\gamma \equiv \left\{ \sum_{i=1}^N \alpha_i k(x_i, \cdot) : x_1, \dots, x_N \in \mathbb{R}^d, \alpha_1, \dots, \alpha_N \in \mathbb{R}, N \in \mathbb{N} \right\}$$

### Theorem

$\forall \gamma > 0 \forall f : \mathbb{R}^d \rightarrow \mathbb{R}$  continuous,  $\forall \varepsilon > 0$

$\exists g \in H_\gamma$  that approximates  $f$  with error  $\varepsilon$

We define a function with  $H$ . We see the  $\cdot$  and this tell us is a function. So we can evaluate every kind of  $x$  point in  $\cdot$  position.

We are able to get a super parametric algorithm and transform it in a non-parametric algorithm. Parametric algorithms is defined by an arbitrary number of parameter we cannot adapt it for every case.

Gaussian Kernels enable consistency by using feature expansion with infinitely many components.

# Lecture 19 - 18-05-2020

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \phi : X \rightarrow H$$

where  $X \rightarrow \mathbb{R}^2$  and  $H \rightarrow \mathbb{R}^N$

$$H_\delta = \left\{ \sum_{i=1}^N \alpha_i k_\delta(x_i, \cdot), x_1, \dots, x_N \in \mathbb{R}^d, \alpha_1, \dots, \alpha_N, N \in \mathbb{N} \right\}$$

Inner product measures "similarities" between data points.

$$\langle x, x' \rangle = \|x\| \|x'\| \cos \Theta \quad x \in X \quad k(x, x')$$

$k$  says how much similar are the structure (tree, documents etc).

I would like to learn a predictor based on the notion of similarity.

$$k(x, x') = \langle \phi(x), \phi(x') \rangle$$

where  $\langle \rangle$  is the inner product.

So we have Data  $\rightarrow$  Kernel  $\rightarrow$  Kernel learning Algorithm

Kernels offer a uniform interface to data in such way they algorithm can learn from data.

Given  $K$  on  $X$ , I need to find  $\exists H_k \quad \phi_k : X \rightarrow H_k$

$\exists \langle \dots \rangle_k$  s.t  $k(x, x') = \langle \phi_k(x), \phi_k(x') \rangle_k$

## Theorem

Given  $K : X \times X \rightarrow \mathbb{R}$ , symmetric

Then  $K$  is a Kernel iff  $\forall m \in \mathbb{N} \forall x_1, \dots, x_m \in X$

The  $m \times m$  matrix  $K \quad K_{ij} = k(x_i, x_j)$  is positive semidefinite

$\forall \alpha \in \mathbb{R}^m \quad \alpha^T K \alpha \geq 0$

In general, given a Kernel  $K$  there is not unique representation for  $\phi_k$  and  $\langle \dots \rangle_k$  (inner product).

However, there is a "canonical" representation:  $\phi_k(x) = K(x, \cdot)$

$$\phi_k : X \rightarrow H \quad H_k = \left\{ \sum_{i=1}^N \alpha_i k(x_i, \cdot), \alpha_1, \dots, \alpha_N \in \mathbb{R}, x_1, \dots, x_N \in X, N \in \mathbb{N} \right\}$$

We have to define an inner product like:

$$\langle \phi_k(x), \phi_k(x') \rangle_k = k(x, x')$$

This is the canonical representation that helps mapping.

What happens to use this mechanism to perform predictions?

$$x \in \mathbb{R}^d \quad w \in \mathbb{R}^d \quad w^T x \quad \text{where } g = \sum_{i=1}^N \alpha_i k(x_i, \cdot)$$

$$\phi_k(x) \quad g \in H_k \quad \langle g, \phi_k(x) \rangle_k = \left\langle \sum_i \alpha_i k(x_i, \cdot), \phi_k(x) \right\rangle_k =$$

We have to satisfy allinearity

$$= \sum_i \alpha_i \langle k(x_i, \cdot), k(x, \cdot) \rangle_k = \sum_i \alpha_i \langle \phi(x_i), \phi_k(x) \rangle_k = \sum_i \alpha_i k(x_i, x) = g(x)$$

At the end we have:

$$\langle g, \phi_k(x) \rangle_k = g(x)$$

Now, if I have two functions:

$$f = \sum_{i=1}^N \alpha_i k(x_i, \cdot) \quad g = \sum_{j=1}^M \beta_j k(x'_j, \cdot) \quad f, g \in H_k$$

$$\langle f, g \rangle_k = \left\langle \sum_i \alpha_i k(x_i, \cdot), \sum_j \beta_j k(x'_j, \cdot) \right\rangle_k = \sum_i \sum_j \alpha_i \beta_j \langle k(x_i, \cdot), k(x'_j, \cdot) \rangle_k =$$

$$= \sum_i \sum_j \alpha_i \beta_j k(x_i, x'_j)$$

$$\|f\|^2 = \langle f, f \rangle_k = \sum_{ij} \alpha_i \alpha_j k(x_i, x_j)$$

Perceptron convergence theorem in kernel space:

$$M \leq \|U\|^2 (\max_t \|x_t\|^2) \quad \forall u \in \mathbb{R}^d \quad y_t u^T x_t \geq 1 \quad \forall g \in H_k \quad y_t g(x_t) \geq 1$$

we know that:

$$\|x_t\|^2 \rightsquigarrow \|\phi_k(x_t)\|_k^2 = \langle \phi_k(x_t), \phi_k(x_t) \rangle_k = k(x_t, x_t)$$

so

.... MANCA ULTIMA FORUMA

Ridge regression:

$$w = (\alpha I + S^T S)^{-1} S^T y$$

$S$  is  $m \times d$  matrix whose rows are the training points  $x_1, \dots, x_m \in \mathbb{R}^d$   
 $y = (y_1, \dots, y_m)$   $y_t \in \mathbb{R}^d$  training labels  $\alpha > 0$

$$(\alpha I + S^T S)^{-1} S^T = S^T (\alpha I_m + S S^T)^{-1}$$

where  $d \times d$  and  $d \times m = d \times m$  and  $m \times m$

$$(S S^T)_{ij} = x_i^T x_j \quad \rightsquigarrow \langle \phi(x_i), \phi(x_j) \rangle_k = k(x_i, x_j) = K_{ij}$$

$$S^T = [x_1, \dots, x_m] \rightsquigarrow [\phi_k(x_1), \dots, \phi_k(x_m)] = [k(x_1, \cdot), \dots, k(x_m, \cdot)] = k(\cdot)$$

$$k(\cdot)^T (\alpha I_m + K)^{-1} y = g$$

where  $1 \times m$  and  $m \times m$  and  $m \times 1$

How to compute prediction?

$$g(x) = y^T (\alpha I_m + K)^{-1} k(x)$$

$1 \times m$  and  $m \times m$  and  $m \times 1$

In fact, is the evaluation of  $g$  in any point  $x$ .

The drawback is that we pass from  $d \times d$  matrix to a  $m \times m$  matrix that can be huge. So it is not really efficient in this way, we need to use additional "tricks" having a more compact representation of the last matrix prediction.



## 19.1 Support Vector Machine (SVM)

It is a linear predictor and is a very popular one because has better performance than perceptron and we will see it for classification but there are also version for regression.

The idea here is that you want to come up with an hyperplane that is defined as a solution of an optimisation problem.

We have a classification dataset  $(x_1, y_1) \dots (x_m, y_m)$   $x_t \in \mathbb{R}^d$   $y_t \in \{-1, 1\}$  and it is linearly separable.

Sum as the solution  $w^*$  (optimisation problem) to this problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|^2 \quad s.t \quad y_t w^T x_t \geq 1 \quad t = 1, 2, \dots, m$$

Geometrically  $w^*$  corresponds to the maximum margin separating hyperplane like:

$$\gamma^* = \max_{u: \|u\|=1} y_t u^T x_t \quad t = 1, \dots, m$$

$u^*$  is achieving  $\gamma^*$  is the maximal margin separator.

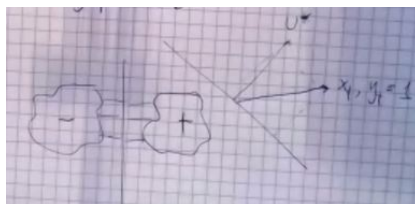


Figure 19.1: Draw of SVG

So I want to maximise this distance.

$$\max_{\gamma > 0} \gamma^2 \quad s.t \quad \|u\|^2 = 1 \quad y_t u^T x_t \geq \gamma \quad t = 1, \dots, m$$

So we can maximise instead of minimising.

What is the theorem? The equivalent between this two.

### Theorem:

$\forall$  linear separator  $(x_1, y_1) \dots (x_m, y_m)$

The max margin separator  $u^*$  satisfies  $u^* = \gamma^* w^*$  where  $w^*$  is the SVM solution and  $\gamma^*$  is the maximum margin.

# Lecture 20 - 19-05-2020

# Bibliography