Statistical Methods for Machine Learning

Andrea Ierardi
Data Science and Economcis
Università degli Studi di Milano

April 12, 2020

Abstract

This is the paper's abstract ...

1 Lecture 1 - 09-03-2020

1.1 Introduction

This is time for all good men to come to the aid of their party!

MACHINE LEARNING 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. 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 —¿ 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 —¿ 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 were 0 is not spam and 1 is spam and that 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. Spam filtering We have two main mistakes: 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. Even in binary classification, mistakes are not equal. e Iotf.TFprIuos.uos True came razee Cussler aircN TASK spam ACG FIRM ftp.y GO IF F Y n is soon IF FEY 0 Nor spam ZERO CNE Cass n n Span No Seamy Binary Classification I 2 FALSE PEENE Mistake Y NON SPAM J Spam FN Mistake i f SPAM y NO spam 2 IF Fp Meter Airenita f Y F on positive y ye en MISTAKE 0 otherwise

Outline The remainder of this article is organized as follows. Section 11 gives account of previous work. Our new and exciting results are described in Section 12. Finally, Section 13 gives the conclusions.

2 Lecture 2 - 07-04-2020

2.1 Argomento

Classification tasks
Semantic label space Y
Categorization Y finite and
small Regression Y appartiene ad —R
How to predict labels?
Using the lost function → ..
Binary classification
Label space is Y = -1, +1
Zero-one loss

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

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

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

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

— DISEGNO —

Some inconvenient properties:

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

2.2.2 Square Loss

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

– DISEGNO —
Derivative :

- more informative
- and differentible

Real numbers as label → 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

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

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

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

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

$y \in Y$ $\hat{y} \in Z$
quadHow 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.

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

If i want to expand the punishment i use logarithmic loss

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

$F(\hat{y}) \to$ *can be 0 if i predict with certainty*
If $\hat{y} = 0.5 \qquad \ell(y, \frac{1}{2}) = ln2$ *constant losses in each prediction*

$\lim_{\hat{y}\to 0^+} \ell(1, \hat{y}) = +\infty$
We give a vanishing probability not rain but tomorrow will rain.
So this is $+\infty$
$\lim_{\hat{y}\to 1^-} \ell(0, \hat{y}) = +\infty$

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

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

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 vec-

tors.
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 $x = (x_1, .., x_d)$
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 xi = 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 $f : X \longrightarrow Y$ (in weather $X \longrightarrow Z$
2. X is our data space (where points live)
3. $X = \mathbb{R}^d$ images
4. $X = X_1 x ... x X_d$ 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 and 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.

3 Lecture 3 - 07-04-2020

4 Lecture 4 - 07-04-2020

5 Lecture 5 - 07-04-2020

6 Lecture 6 - 07-04-2020

7 Lecture 7 - 07-04-2020

8 Lecture 8 - 07-04-2020

9 Lecture 9 - 07-04-2020

10 Lecture 10 - 07-04-2020

10.1 TO BE DEFINE

$|E[z] = |E[|E[z|x]]$

$|E[X] = \sum_{t=1}^{m} |E[x\Pi(At)]$

$x \in \mathbb{R}^d$
$\mathbb{P}(Y_{\Pi(s,x)} = 1) =$

$\mathbb{E}[\Pi Y_{\Pi(s,x)} = 1] =$

$= \sum_{t=1}^{m} \mathbb{E}[\Pi\{Y_t = 1\} \cdot \Pi Pi(s,x) = t] =$

$= \sum_{t=1}^{m} \mathbb{E}[\mathbb{E}[\Pi\{Y_t = 1\} \cdot \Pi\{\Pi(s,x) = t\}|X_t]] =$

$given the fact that Y_t \sim \eta(X_t) \Rightarrow give me probability$
$Y_t = 1 and \Pi(s,x) = t are independent given X_Y (e.g. \mathbb{E}[Zx] = \mathbb{E}[x] * \cdot \mathbb{E}[z]$

$= \sum_{t=1}^{m} \mathbb{E}[\mathbb{E}[\Pi\{Y_t = 1\}|X_t] \cdot \mathbb{E}[\Pi(s,x) = t|Xt]] =$

$= \sum_{t=1}^{m} \mathbb{E}[\eta(X_t) \cdot \Pi \cdot \{\Pi(s,x) = t\}] =$

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

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

$\mathbb{P}(Y_{\Pi(s,x)} = 1, y = -1) =$

$= \mathbb{E}[\Pi\{Y_{\Pi(s,x)} = 1\}\dot{\Pi}\{Y = -1|X\}]] =$

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

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

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

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

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

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

$\mathbb{E}[\ell_D(\hat{h}_s)] = \mathbb{P}(Y_{\Pi(s,x)} \neq y) =$

$= \mathbb{P}(Y_{\Pi(s,x)} = 1, y = -1) + \mathbb{P}(Y_{Pi(s,x)} = -1, y = 1) =$

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

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

MANCAAAAAAA ROBAAA

$\eta(x') <= \eta(x) + c||X - x'|| - - > euclideandistance$

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

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

$\eta(X) \cdot (1 - \eta(x')) + (1 - \eta(x)) \cdot \eta(x') <=$

$<= \eta(x) \cdot ((1 - \eta(x)) + \eta(x) \cdot c||X - x'|| + (1 - \eta(x)) \cdot c||X - x'|| =$

$= 2 \cdot \eta(x) \cdot (1 - \eta(x)) + c||X - x'||$

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

Compare risk for zero-one loss
$\mathbb{E}[min\{\eta(x), 1 - \eta(x)\}] = \ell_D(f*)$

$\eta(x) \cdot (1 - \eta(X)) <= min\{\eta(x), 1 - eta(x)\} \quad \forall x$

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

$\mathbb{E}[\ell_d(\hat{l}_s)] <= 2 \cdot \ell_D(f*) + c \cdot \mathbb{E}[||X - X_{\Pi(s,x)}||]$

$\eta(x) \in \{0, 1\}$

Depends on dimension: curse of dimensionality

–DISEGNO–

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