

# Lecture 7 - 07-04-2020

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**

## 1.1 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.

## 1.2 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

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

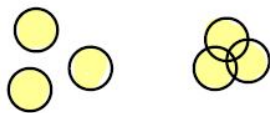


Figure 1.1: Example

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 1.2: 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.

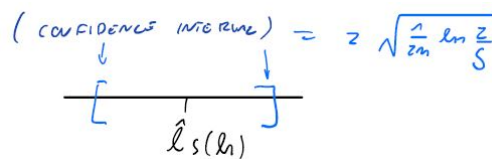


Figure 1.3: Example

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  $h^*$**

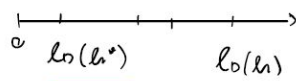


Figure 1.4: Example

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

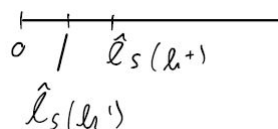


Figure 1.5: Example

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

$$\begin{aligned} f^* \text{ is } \mathbf{Bayes Optimal} \text{ for } (D, \ell) \\ \forall h \quad \ell_D(h) \geq \ell_D(f^*) \end{aligned}$$

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

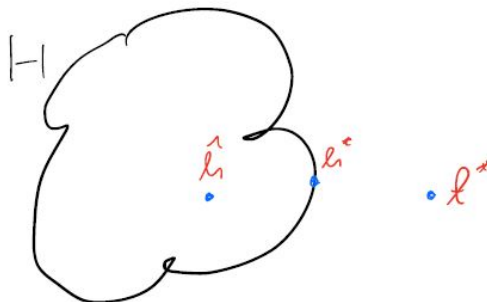


Figure 1.6: Draw of how  $\hat{h}$ ,  $h^*$  and  $f^*$  are represented

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

### 1.3 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

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

$$\begin{aligned} \text{We know } \ell_D(\hat{h}) - \ell_D(h^*) &\leq 2 \cdot \max |\hat{\ell}_S(h) - \ell_D(h)| \quad \Rightarrow \\ &\Rightarrow \exists h \in H \quad |\hat{\ell}_S(h) - \ell_D(h)| > \frac{\varepsilon}{2} \quad \Rightarrow \end{aligned}$$

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 ( $\mathbb{P}(\dots)$ )**

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

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

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

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

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