EE-411: Fundamentals of inference and learning **2021-2022**

Lecture 8: A short course on statistical learning theory

Lecturer: Florent Krzakala *Scribes: Santiago Gallego Restrepo, Jad Sobhie,*

8.1 Intro

When performing a supervised learning task, one trains the model such that it minimizes the empirical risk. However, what we would like is for the population risk to minimized.

Definition 8.1 (Population Risk) *Let $\mathbf{x} \in \mathcal{X}$ (population set), y be the true label and $\hat{y}$ be the estimator,*

$$\mathcal{R}_{pop}[\hat{y}] := \mathbb{E}_{\mathbf{x},y}[loss(\hat{y}(\mathbf{x}), y)]$$

Definition 8.2 (Empirical Risk) *Let n be the number of observations,*

$$\mathcal{R}^{(n)}_{emp}[\hat{y}] := \frac{1}{n}\sum_{i=1}^{n} loss(\hat{y}(\mathbf{x}_i), y_i)$$

By the LLN, we expect for any function $\hat{y}$

$$\mathcal{R}^{(n)}_{\text{emp}}[\hat{y}] \xrightarrow{p} \mathcal{R}_{\text{pop}}[\hat{y}]$$

Therefore, given a "large" dataset, minimizing the empirical risk should suffice (i.e. population risk is minimized as well).

The question now is, how large should n be?

This is the question that statictical learning theory aims to answer. More precisely, it aims to give some insight on the difference between $\mathcal{R}_{\text{pop}}[\hat{y}]$ and $\mathcal{R}^{(n)}_{\text{emp}}[\hat{y}]$.

For concreteness, we will be using using as loss function, the 0/1 loss:

$$\begin{aligned}\text{loss}(\hat{y}(\mathbf{x}), y) &:= 1 - \mathbf{1}_{\hat{y}(\mathbf{x})=y} \\ &= \begin{cases} 0, & \text{if } \hat{y}(\mathbf{x}) = y \\ 1, & \text{otherwise} \end{cases}\end{aligned}$$

8.2 For a given function $\hat{y}()$

Theorem 8.3 (Hoeffding bound Reminder) *Let $Z_i = 0/1$ a binary variable with $i = 1, ...n$. Then,*

$$\mathbb{P}\left(\frac{1}{n}\sum_i Z_i \leq \mathbb{E}[Z] - \epsilon\right) \leq \delta \tag{8.1}$$

with $\delta = e^{-2n\epsilon^2}$.

Developing this inequality , we get :

$$1 - \mathbb{P}\left(\frac{1}{n}\sum_i Z_i \leq \mathbb{E}[Z] - \epsilon\right) \geq 1 - \delta$$

$$\mathbb{P}\left(\frac{1}{n}\sum_i Z_i \geq \mathbb{E}[Z] - \epsilon\right) \geq 1 - \delta$$

$$\mathbb{P}\left(\mathbb{E}[Z] - \frac{1}{n}\sum_i Z_i \leq \epsilon\right) \geq 1 - \delta$$

This means that for a random binary variable and for a probability at least δ ,the difference between the true mean and the new estimate of the mean is smaller than ϵ .
Another way to formulate this result would be to write

$$\delta = e^{-2n\epsilon^2} \;\rightarrow\; \log\frac{1}{\delta} = 2n\epsilon^2 \;\rightarrow\; \epsilon = \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

So if the probability $\mathbb{P}$ is at least $1 - \delta$, then

$$\mathbb{E}[Z] - \frac{1}{n}\sum_i Z_i \leq \sqrt{\frac{1}{2n}\log\frac{1}{\delta}} \approx \sqrt{\frac{1}{n}}$$

for a small δ according to PAC learning

R This result is only true for a given function of a specific parameter and can't be applied to a class of functions especially when the function we are looking for is the one that minimizes the empirical risk .

8.3 Finite number of functions

Definition 8.4 (Uniform Convergence) *Let $\hat{y}_1, ..., \hat{y}_N$ be N estimators, n be the number of observations (i.e. datapoints) and $\epsilon > 0$.*

Then, with probability of at least $1 - \delta$ (with $\delta := e^{-2n\cdot\epsilon^2}$), we have

1. $\forall \hat{y} \in \mathcal{F}, \quad \mathcal{R}_{pop}[\hat{y}] - \mathcal{R}^{(n)}_{emp}[\hat{y}] \leq \sqrt{\frac{\log(N)+\log(1/\delta)}{2n}}$
2. *Equivalently,* $\underset{y\in F}{Sup}\left\{[\mathcal{R}_{pop}[\hat{y}] - \mathcal{R}^{(n)}_{emp}[\hat{y}]\right\} \leq \sqrt{\frac{\log(N)+\log(1/\delta)}{2n}}$

R If $n >> \log N$, then "we're good" (i.e. $\mathcal{R}_{\text{pop}}[\hat{y}] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}]$ is bounded). In other words, if we have more datapoints than functions, then $\mathcal{R}_{\text{pop}}[\hat{y}] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}]$ should be bounded.

Proof:

$$\mathbb{P}\Big[\forall k \in \{1, .., N\}, \quad \mathbb{E}_{\mathbf{x},y}[\text{loss}(\hat{y}_k(\mathbf{x}), y)] - \frac{1}{n}\sum_{i=1}^{n}\text{loss}(\hat{y}_k(\mathbf{x}_i), y_i) \leq \epsilon\Big] = 1 - \mathbb{P}\Big[\exists k \in \{1, .., N\}, \quad \mathcal{R}_{\text{pop}}[\hat{y}_k] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}_k] \geq \epsilon\Big]$$

What's more, we have that

$$\mathbb{P}\Big[\exists k \in \{1,..,N\}, \mathcal{R}_{\text{pop}}[\hat{y}_k] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}_k]\Big] \leq \sum_{j=1}^{N} \mathbb{P}[\mathcal{R}_{\text{pop}}[\hat{y}_j] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}_j] \geq \epsilon] \leq N \cdot e^{-2n\cdot\epsilon^2}$$

R The inequality on the left is obtained from Boole's inequality. As for the one on the right, it is given by Hoeffding's inequality

Which implies,

$$\mathbb{P}\Big[\forall k \in \{1,..,N\}, \quad \mathbb{E}_{\mathbf{x},y}[\text{loss}(\hat{y}_k(\mathbf{x}), y)] - \frac{1}{n}\sum_{i=1}^{n} \text{loss}(\hat{y}_k(\mathbf{x}_i), y_i) \leq \epsilon\Big] \geq 1 - N \cdot e^{-2n\epsilon^2}$$

■

8.4 What to do if N is infinite

Definition 8.5 (Growth Function) *Let $\hat{y} \in \mathcal{F}$. We introduce $\Pi_{\mathcal{F}}\Big[\vec{x_1}, ..., \vec{x_m}\Big]$ as the number of possible classifications.*

$$\Pi_{\mathcal{F}}\Big[\vec{x_1}, ..., \vec{x_m}\Big] \leq 2^m$$

We define the Growth Function:

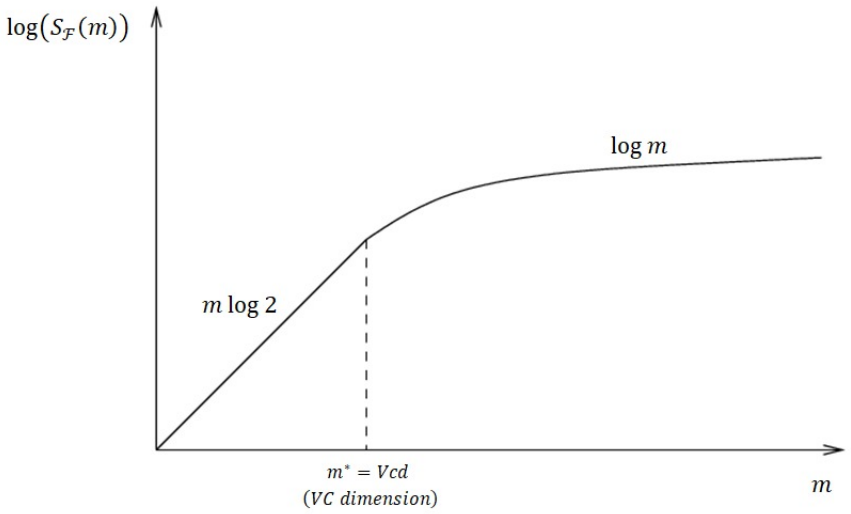
$$S_{\mathcal{F}}(m) = MAX \text{ all positions of } m \text{ points } \Big[\Pi_{\mathcal{F}}[\vec{x_1}, ..., \vec{x_m}]\Big] \leq 2^m$$

R *The Growth Function is a measure of complexity of the function class (a bigger Growth Function value translates into the function class being able to fit more complicated labels).*

Using the works of Vapnik–Chervonenkis (VC), we can prove that if $\hat{y} \in \mathcal{F}$ with probability of at least $1-\delta$, then:

$$\underset{\hat{y}}{\text{Sup}} \Big\{[\mathcal{R}_{\text{pop}}[\hat{y}] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}]\Big\} \leq 2\sqrt{\frac{\log(S_{\mathcal{F}}(m)) + \frac{4}{\delta}}{n}}$$

Looking further into the (log of the) Growth Function, we can aslo notice that its typical behavior goes as follows:



This shows us that for any function class, there exists a point m^* beyond which we wouldn't be able to fit a possible separation.

Theorem 8.6 (Vapnik–Chervonen(VC), Sauer–Shelah, Sauer's lemma)

$$S_{\mathcal{F}}(m) < \left[\frac{em}{Vcd}\right]^{Vcd}$$

With $Vcd = \max_m$ *such that* $S_{\mathcal{F}}(m) = 2^m$

(More intuitively, Vcd is the maximum number of points that the function class can fit perfectly, regardless of position or label.)

R *If* $\mathcal{F}$ *is the set of linear function in dimension* d*, then* $Vcd(\mathcal{F}) = d + 1$

Definition 8.7 (VC Bound) *Let* $\hat{y} \in \mathcal{F}$*, then with probability of at least* $1 - \delta$*:*

$$\sup_{\hat{y}} \left\{ [\mathcal{R}_{pop}[\hat{y}] - \mathcal{R}^{(n)}_{emp}[\hat{y}] \right\} \leq 2\sqrt{\frac{2Vcd\log(\frac{2ne}{Vcd}) + \log(\frac{4}{\delta})}{n}} \leq \sqrt{\frac{Vcd\log(n)}{n}k}$$

Therefore, if $n >> Vcd$ *then* $\mathcal{R}_{pop}[\hat{y}] - \mathcal{R}^{(n)}_{emp}[\hat{y}]$ *is very small, allowing us to focus on optimising the empirical risk instead of the population risk .*

8.5 More on Uniform Convergence

Let $f \in \mathcal{F}$ (a class of function) and $\text{VCd}(\mathcal{F}) = $ VC dimension of the function class. Then, from the uniform convergence theorem, we have

$$\sup_{f \in F} |\mathcal{R}_{\text{pop}}[\hat{y}] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}]| \lesssim \sqrt{\frac{\text{VCd}(\mathcal{F}) \log n}{n}}$$

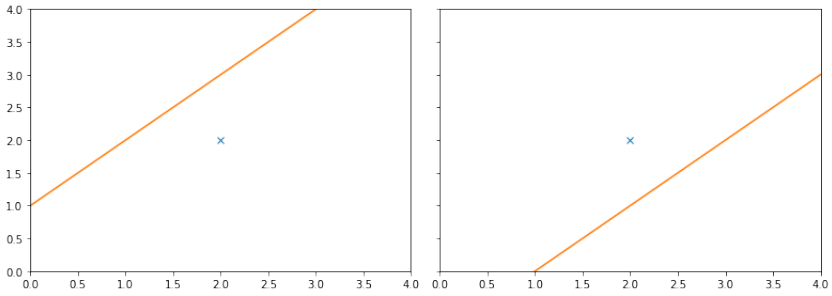
R It is possible to prove that $\sup_{f \in F} |\mathcal{R}_{\text{pop}}[\hat{y}] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}]| \lesssim \sqrt{\frac{\text{VCd}(\mathcal{F})}{n}}$. The proof however, is not trivial.

The question now is, what is $\text{VCd}(\mathcal{F})$?

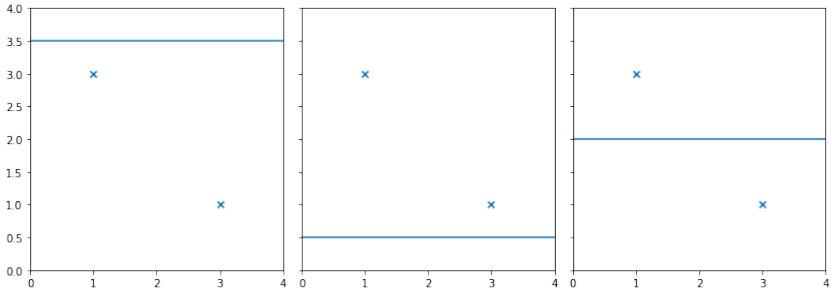
Definition 8.8 (Vapnik–Chervonenkis dimension) *$VCd(\mathcal{F})$ is the largest number of points, such that it exists a set of points that can be fitted, no matter the assignment $f \in \mathcal{F}$.*

8.5.1 Example 1: Linear separation in dimension $d = 2$

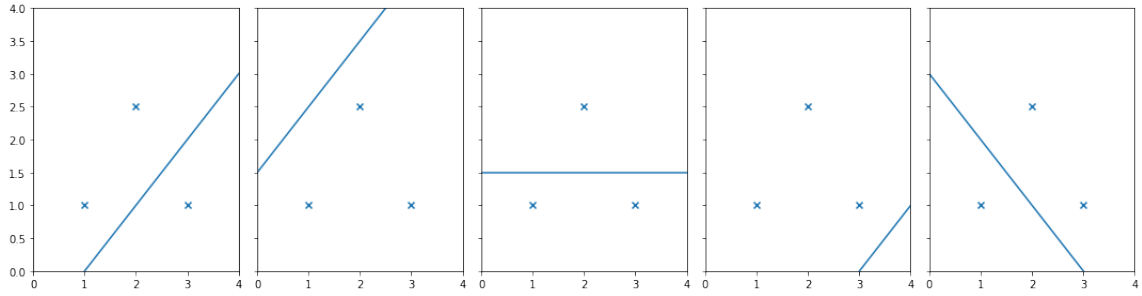
$n = 1$



$n = 2$



$n = 3$



For $n = 3$, any set of unaligned points can be fitted, no matter the assignment $f \in \mathcal{F}$. On the contrary, for $n = 4$ there isn't a set of points that can be fitted no matter the assignment.

Therefore, VCd[linear - class, $d = 2$] $= 3$.

Claim 8.9 *VCd[linear -class, d]* $= d + 1$

Conclusion: For linear classifiers, we have $\sup_{f \in F} |\mathcal{R}_{\text{pop}}[\hat{y}] - \mathcal{R}^{(n)}_{\text{emp}}[\hat{y}]| \lesssim \sqrt{\frac{d+1}{n}}$