Foundations of Machine Learning

Paris Dauphine - PSL
Lecturer: Yann Chevaleyre
Scribe: Jules Merigot

Lecture 4

October 19, 2023

Uniform Convergence and Rademacher Complexity

RECAP from Previous Lecture (Liva Ralaivola, Criteo)

Hoeffding Inequality

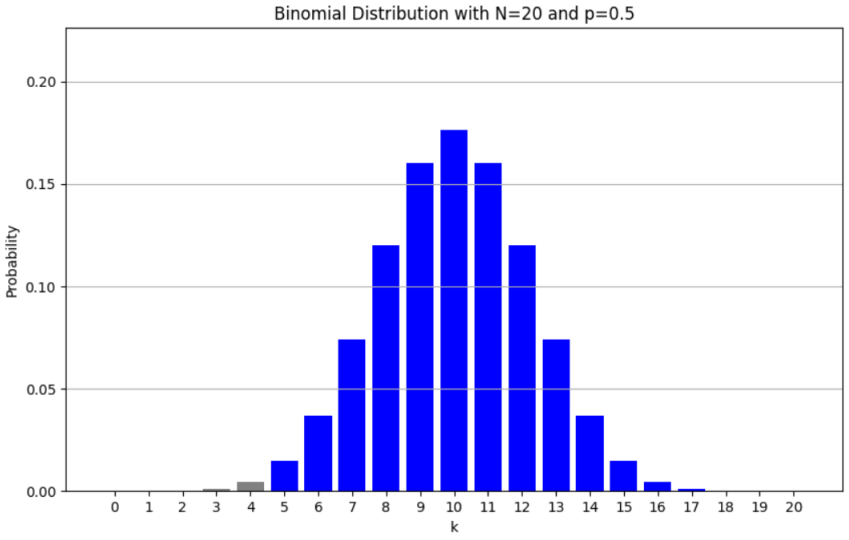
Let $Z_1, ..., Z_N \in [0, 1]$ be independently and identically distributed random variables, then:

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}[Z] \geq \epsilon\right) \leq exp(-2N\epsilon^2)$$

$$\iff \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}[Z]\right| \geq \epsilon\right) \leq 2\,exp(-2N\epsilon^2)$$

Following graphs are from this Jupyter notebook: Hoeffding notebook

Example: We have a Binomial distribution with N = 20, p = 0.5:



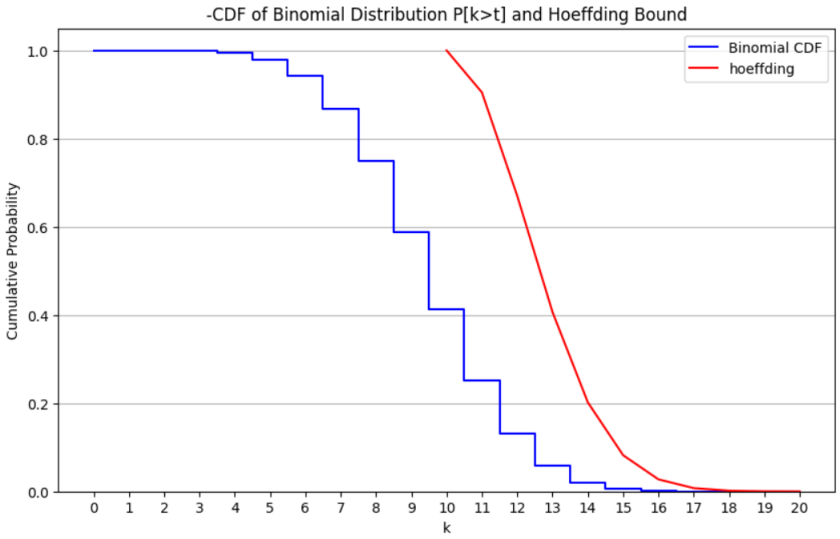
Area to the right of threshold t: $\mathbb{P}(k > 5) = 0.9941$

Hoeffding bound: $\mathbb{P}(k > 5) < 0.0821$

Hoeffding formula applied here:

$$\mathbb{P}(k \geq t) \leq exp(-2N(\frac{t}{20 - 0.5})^2)$$

CDF of Binomial Distributiom $\mathbb{P}[k > t]$ and Hoeffing Bound



Exercise: How can we go from the first Hoeffding Inequality

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}[Z] \geq \epsilon\right) \leq exp(-2N\epsilon^2)$$

to this Hoeffding formula

$$\mathbb{P}(k \geq t) \leq exp(-2N(\frac{t}{20-0.5})^2)$$

PROOF:

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}[Z] \geq \epsilon\right) \leq exp(-2N\epsilon^2)$$

$$\Rightarrow \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i \geq \epsilon + \mathbb{E}[Z]\right) \leq exp(-2N\epsilon^2)$$

$$\Rightarrow \mathbb{P}\left(\sum_{i=1}^{N} Z_i \geq N\epsilon + N\,\mathbb{E}[Z]\right) \leq exp(-2N\epsilon^2)$$

Here, we can define the following:

$$k = \sum_{i=1}^{N} Z_i\,, \qquad t = N\epsilon + N\,\mathbb{E}[Z]\,, \qquad \epsilon = \frac{t}{N} - \mathbb{E}[Z]$$

Therefore,

$$\Rightarrow \mathbb{P}(k \geq t) \leq exp(-2N(\frac{t}{N} - \mathbb{E}[Z])^2)$$

1 Introduction

Reminder on Hoeffding Inequality:

A sequence of random variables $\{Z_1, Z_2, \ldots, Z_n\}$ converges in probability to Z as N approaches ∞ if and only if:

$$\forall\, \epsilon, \delta \ \in]0,1[\ ,\ \exists n, \quad \text{if } N > n,$$
$$|Z_N - Z| < \epsilon \quad \text{with probability at least } 1-\delta$$

Equivalently, $\forall\, \epsilon, \delta \ \in]0,1[$, there exists a function $n(\epsilon, \delta)$, such that:

$$\forall\, \epsilon, \delta \ \in]0,1[\ ,\ N > n(\epsilon,\delta) \ \Rightarrow \ |Z_N - Z| < \epsilon \text{ with probability at least } 1-\delta$$

Exercise: Show that in the Hoeffding setting,

$$\frac{1}{N}\sum_{i=1}^{N} Z_i \xrightarrow[N\to\infty]{\text{in prob.}} \mathbb{E}(Z), \quad \text{and give } n(\epsilon,\delta).$$

ANSWER:

First we can see that:

$$\mathbb{E}[Y_N] \ = \ \frac{1}{N}\sum \mathbb{E}[Z_i] \ = \ \mathbb{E}[Z]$$

Therefore, we can rewrite Hoeffding as:

$$P(|Y_N - \mathbb{E}[Z]| > \epsilon) \ < \ 2e^{-2N\epsilon^2} \ \leq \ \delta$$

Now we can solve for N:

$$\Rightarrow \ 2e^{-2N\epsilon^2} \leq \delta \iff -2N\epsilon^2 \leq log(\frac{\delta}{2}) \iff N \geq \frac{log(\frac{\delta}{2})}{2\epsilon^2}$$

$$\text{Therefore,} \quad n(\epsilon,\delta) = \frac{log(\frac{\delta}{2})}{2\epsilon^2}, \quad \text{so now we know that if} \quad N > n(\epsilon,\delta),$$
$$\text{then } \Big|Y_N - \mathbb{E}[X]\Big| \leq \epsilon \ \text{, with probability at least } 1-\delta$$

Hoeffding Inequality (Another form):

$$\iff \ \mathbb{P}\Big(\Big|Y_N - \mathbb{E}[Z]\Big| > \epsilon\Big) \ < \ \delta \ \text{ for } \ 2e^{-2N\epsilon^2} \leq \delta$$

$$\text{so,} \quad 2N\epsilon^2 \leq log(\frac{2}{\delta}) \ \Rightarrow \ \epsilon \leq \sqrt{\frac{log(\frac{2}{\delta})}{2N}}$$

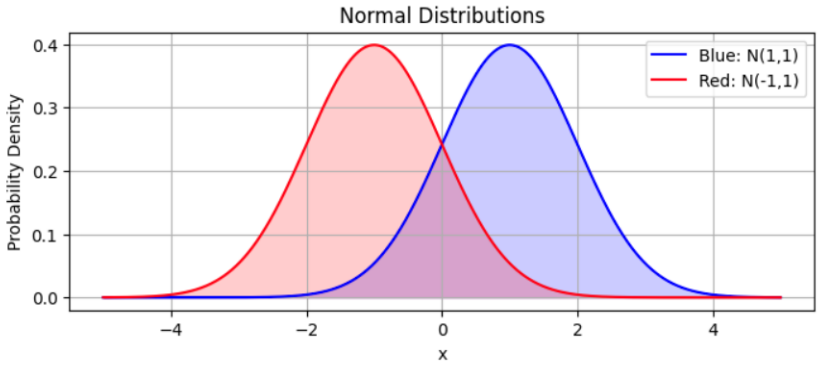
$$\iff \boxed{\Big|\frac{1}{N}\sum Z_i - \mathbb{E}[Z]\Big| < \sqrt{\frac{log(\frac{2}{\delta})}{2N}} \text{ with probability at least } 1-\delta}$$

Very useful form, and will be used often

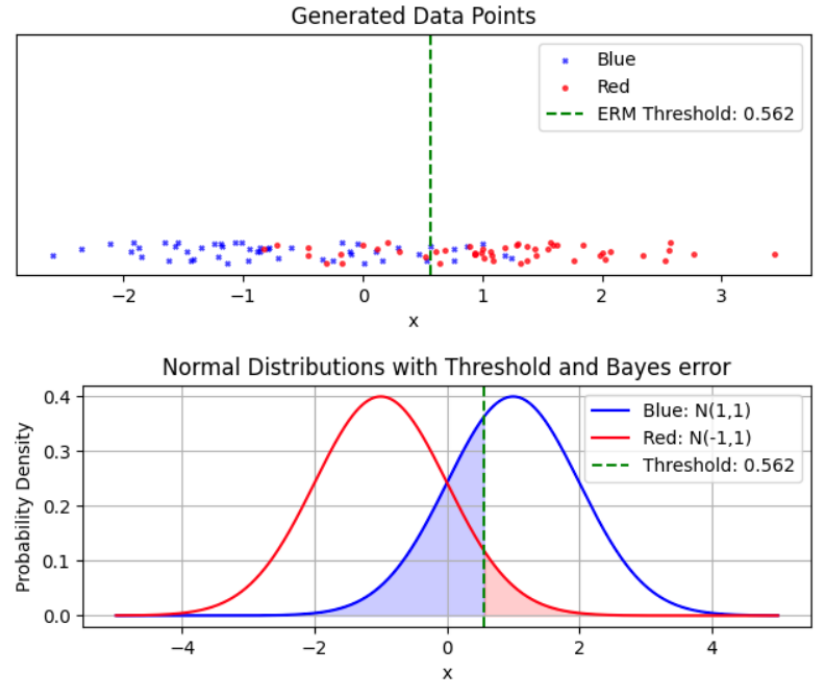
ERM - Expected Empirical Risk

Following graphs are from this Jupyter notebook: Two Gaussians notebook

Bayes Risk:



Bayes risk computed is in the area where the two normal distributions overlap.



$\hat{R}(f_{ERM}) = 0.180$
$R(f_{ERM}) = 0.195$

How can we see $\hat{R}(f_{ERM})$ as a random variable?
Start by bounding $\hat{R}(0) - R(0)$:

$$\mathbb{P}_{S\sim\mathcal{P}^N}[\hat{R}_S(f_0) > R(f_0)]$$

where S is the data, $\hat{R}_S$ is the empirical risk, f_0 is the threshold classifier on 0, and $R(f_0)$ is the true risk

$$\mathbb{P}_{S\sim\mathcal{P}^N}[\hat{R}_S(f_0) > R(f_0)]$$

is bounded between 0 and 1, so we can directly apply Hoeffding's bound.

Applying Hoeffding:

$$\left|\hat{R}_S(f_0) > R(f_0)\right| < \sqrt{\frac{log(\frac{2}{\delta})}{2N}} \text{ with probability at least } 1-\delta$$

Note: $\hat{R}_S(f_0)$ can also be expressed as $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{[f_0(x_i)\neq y_i]}$

This can therefore also be written as:

$$\Rightarrow \quad \mathbb{P}(|\hat{R} - R| < \epsilon) > 1 - \delta$$

But, if we want to be (very) precise, we have:

$$z_i = \mathbb{1}_{[f_0(x_i)\neq y_i]}\ , \qquad \hat{R}_S(f_0) = \frac{1}{N}\sum z_i\ , \qquad R(f_0) = \mathbb{E}[z_i]$$

We have a Bernoulli distribution, so the probability is constant for all x_i's

$$\iff \quad \left|\frac{1}{N}\sum z_i - \mathbb{E}[z]\right| \quad < \quad \sqrt{\frac{log(\frac{2}{\delta})}{2N}}$$

Question: Can we bound $\left|\hat{R}_S(f_{ERM} - R(f_{ERM})\right|$?
The answer is **NO!**

But why not? A good example to explain is the following:

Example: If a professor gives 10 coins to each of his students and tells them to flip them, and there are infinite students, then there will always be at least one student that has only heads. This cannot be bounded due to the nature of infinity.

To visualize this,

The classifier is extremely overfit, so $\left|\hat{R}_S(f_{ERM} - R(f_{ERM})\right|$ is very big.
This is because $\hat{R}_S(f_{ERM}) \to 0$ and $R(f_{ERM}) \to +\infty$.

In this lecture:
- $S = (x_1, y_1), ..., (x_N, y_N)$ is considered a random variable.
- The empirical risk $\hat{R}_S(f_S)$ is defined as:

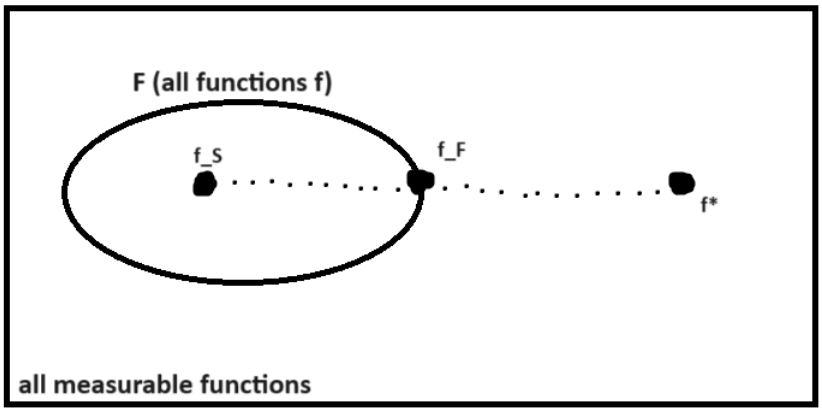
$$\hat{R}_S(f_S) = \frac{1}{N} \sum_{i=1}^{N} l(f_S(x_i), y_i)$$

where $\hat{R}_S(f_S)$ is computed on a dataset if S is present, and $l(.)$ can also be the $0/1$ loss.

2 Notions of Consistency

Recall: A learner f_S is ERM if and only if $\quad f_S \in \arg\min_{f \in \mathcal{F}} \hat{R}_S(f)$

We can use the drawing below to visualize f_S, $f_{\mathcal{F}}$, and $f*$:



Bayes estimator (best estimator): $\quad f* \in \arg\min_{f \in \text{measurable}} R(f)$

Best classifier in the class $\mathcal{F}$: $\quad f_F \in \arg\min_{f \in \mathcal{F}} R(f)$

So, we can say the following for the risks:

$$R(f_S) \quad \geq \quad R(f_{\mathcal{F}}) \quad \geq \quad R(f*)$$

where the difference between $R(f_S)$ and $R(f_{\mathcal{F}})$ is the estimation error, and the difference between $R(f_{\mathcal{F}})$ and $R(f*)$ is the approximation error.

Definition: The learning algorithm f_S is:

- (TOO STRONG) universally Bayes consistent if and only if

$$\forall \text{ distribution } \mathcal{P}, \quad R(f_S) \xrightarrow[N \to \infty]{\text{in prob.}} R(f*)$$

in other words, there is a function $n(\epsilon, \delta, \mathcal{P})$, such that for any $\mathcal{P}, \epsilon, \delta$:

$$\text{if } N > n(\epsilon, \delta, \mathcal{P}) \text{ then,}$$

$$\text{for } S \sim \mathcal{P}^N, \quad |R(f_S) - R(f*)| < \epsilon \text{ with probability } 1 - \delta$$

⚠ This is impossible for ERM

- (TOO WEAK) universally F-consistent if

$$\forall \text{ distribution } \mathcal{P}, \quad R(f_S) \xrightarrow[N\to\infty]{\text{in prob.}} R(f_{\mathcal{F}})$$

- (IN BETWEEN) is a PAC-learner (Probably Approximately Correct) if there is a function $n(\epsilon, \delta)$ such that $\forall$ distribution $\mathcal{P}$

$$\forall \epsilon, \delta \in]0,1[\text{ if } N > n(\epsilon, \delta)$$

$$\text{then for } S \sim \mathcal{P}^N, \quad \left|R(f_S) - R(f_{\mathcal{F}})\right| < \epsilon \text{ with probability } 1-\delta$$

⚠ PAC implies $\mathcal{F}$ consistency

3 PAC Learning and Uniform Convergence for ERM

Reminder: PAC = Probably Approximately Correct

- We want to bound $R(f_S) - R(f_{\mathcal{F}})$
- Hoeffding allows us to bound $R(f) - \hat{R}(f)$ for a fixed f !! (single classifier)

Want to bound:

$$R(f_S) - R(f_{\mathcal{F}}) \;=\; R(f_S) - \hat{R}(f_S) + \hat{R}(f_S) - \hat{R}(f_{\mathcal{F}}) + \hat{R}(f_{\mathcal{F}}) - R(f_{\mathcal{F}})$$

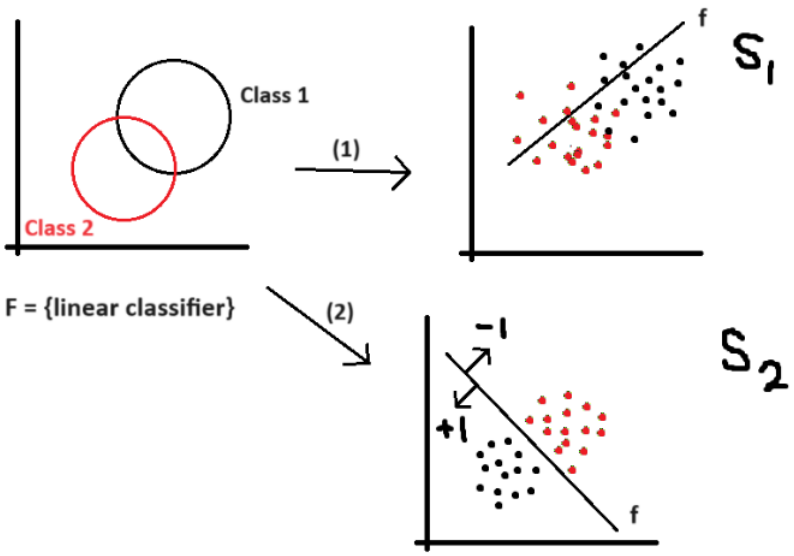
$\hat{R}(f_S)$ is the Risk just on the data, and $\hat{R}(f_{\mathcal{F}})$ is the Risk on all possible functions.

$$R(f_S) - R(f_{\mathcal{F}}) \;\leq\; 2 \sup_{f\in\mathcal{F}} \left|R(f) - \hat{R}(f)\right|$$

sup : *(supremum)* the smallest quantity that is greater than or equal to each of the given set or subset of quantities.

Definition: The unrepresentativeness of S with respect to $\mathcal{F}$ is:

$$\text{UnRep}(\mathcal{F}, S) \;=\; \sup_{f\in\mathcal{F}} \left|R(f) - \hat{R}(f)\right|$$



legend on next page...

$$S_1: \quad \hat{R}(\mathcal{F}) \approx R(f), \text{ so } \mathrm{UnRep}(\mathcal{F}_1, S_1) \ll 1$$
$$S_2: \quad \hat{R}(\mathcal{F}) = 0, \text{ so } \mathrm{UnRep}(\mathcal{F}_1, S_2) \geq 90\%$$

$$(1) \quad S_1 \sim \mathcal{P}^N$$
$$(2) \quad S_2 \sim \mathcal{P}^N$$

Theorem: If, for class $\mathcal{F}$, there exists $n(\epsilon, \delta)$ such that for any distribution $\mathcal{P}$, any $\epsilon, \delta \in]0, 1[$,

$$\text{if } N > n(\epsilon, \delta), \text{ then the } \mathrm{Unrep}(\mathcal{F}, S) < \epsilon \text{ with probability } 1 - \delta$$

(which is called the uniform convergence property)

then, ERM is a PAC learner on $\mathcal{F}$.

Application to finite classes $\mathcal{F}$:

- I want to show that $\mathrm{UnRep}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| < \epsilon$ with probability $1 - \delta$ for $N > n(\epsilon, \delta)$.
- $$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \geq \epsilon) \quad = \quad \mathbb{P}(\exists f \in \mathcal{F}, |R(f) - \hat{R}(f)| \geq \epsilon)$$

$$\boxed{\textbf{Union Bound: } \mathbb{P}(A \cup B) \leq P(A) + P(B) \quad \Rightarrow \quad \mathbb{P}(\exists i, A_i) \leq \textstyle\sum_i \mathbb{P}(A_i)}$$

$$\Rightarrow \quad \mathbb{P}(\exists f \in \mathcal{F}, |R(f) - \hat{R}(f)| \geq \epsilon) \quad \leq \quad \sum_{f \in \mathcal{F}} \mathbb{P}(|R(f) - \hat{R}(f)| \geq \epsilon)$$

(f does not depend on the data $\Rightarrow$ Hoefdding Inequality!)

$$\textbf{Note that:} \quad \hat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i) \,, \quad R(f) = \mathbb{E}_{S \sim P^N}[\hat{R}(f)]$$

Hoeffding tells us:

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| < \epsilon) \quad \leq \quad \sum_{f \in \mathcal{F}} \mathbb{P}(|R(f) - \hat{R}(f)| \geq \epsilon) \quad \leq \quad \delta \times |\mathcal{F}|$$

$$\text{if } N > \frac{log(\frac{2}{\delta}}{2\epsilon^2} \qquad (|\mathcal{F}| \text{ is the cardinality of } \mathcal{F})$$

$\Rightarrow$ $\mathrm{UnRep}(\mathcal{F}, S) \leq \epsilon$ with probability $1 - \delta \times |\mathcal{F}|$ when $N > \frac{log(\frac{2}{\delta}}{2\epsilon^2}$

$\rightarrow$ let $\delta' = \delta \times |\mathcal{F}|$ $\Rightarrow$ $\mathrm{UnRep}(\mathcal{F}, S) \leq \epsilon$ with probability $1 - \delta$ when $N > \frac{log(\frac{2 \times |\mathcal{F}|}{\delta'})}{2\epsilon^2}$

$\Rightarrow$ ERM on finite classes is PAC learnable

Equivalently,

$$\text{UnRep}(\mathcal{F}, S) \leq \frac{log(\frac{2\times|\mathcal{F}|}{\delta'})}{2N} \text{ with probability at least } 1-\delta'$$

$$\Rightarrow \quad |R(f_S) - \hat{R}(f_S)| \leq \frac{2log(\frac{2\times|\mathcal{F}|}{\delta'})}{2N} = \frac{log(\frac{2\times|\mathcal{F}|}{\delta'})}{N} \text{ with probability at least } 1-\delta'$$

⚠ This only works for finite class because the union bound for ∞ number of events looks like $P(\exists i, A_i) \leq \sum_{i=1}^{\infty} P(A_i) \approx \infty$

4 The Case $|\mathcal{F}| = \infty$, Rademacher Complexity

GOAL: Bound $\text{UnRep}(\mathcal{F}, S)$ for $|\mathcal{F}| = \infty$ without using the Union bound (i.e, Neural Networks)

There are **many tools** for this: Vapaik dimension, converging numbers, Gaussian complexity, Rademacher complexity, ...
BUT, Rademacher applies to arbitrary bounded losses, and not just 0/1 losses.

Notation: $Z = (X, Y)$ a labelled example, and $S = (Z_1, ..., Z_N)$.

$$\text{Given } \mathcal{F}, \text{ define } \mathcal{G} = l \circ \mathcal{F} = \{(x, y) \longmapsto l(f(x), y), f \in \mathcal{F}\}$$

$$\begin{aligned}\text{UnRep}(\mathcal{F}, S) &= \sup_{f\in\mathcal{F}} |R(f) - \hat{R}(f)| \\ &= \sup_{g\in\mathcal{G}} \left| \frac{1}{N}\sum_{i=1}^{N} g(Z_i) - \mathbb{E}_{Z\sim P}\, g(Z) \right|\end{aligned}$$

Definition: The empirical Rademacher complexity of S on $\mathcal{G}$ is:

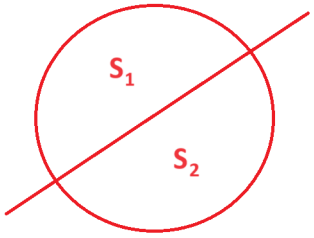
$$\boxed{\hat{Rad}(G) = \frac{1}{N}\,\mathbb{E}_{\sigma_1 ... \sigma_N \sim \text{Unif}(\{-1,1\})} \sup_{g\in\mathcal{G}} \sum_{i=1}^{N} \sigma_i\, g(Z_i)}$$

Intuition 1: Suppose I have drawn 2 datasets S_1, S_2, then:

$$\begin{aligned}\sup_g |\hat{R_{S_1}}(f) - \hat{R_{S_2}}(f)| &= \sup_{g\in\mathcal{G}} \frac{1}{N}\left[\sum_{(x,y)\in S_1} g(Z_1) - \sum_{(x,y)\in S_2} g(Z_2)\right] \\ &= \sup_g \frac{1}{N} \sum_{(x,y)\in S_1\cup S_2} \sigma_i\, g(Z_i) \\ &\text{where } \sigma_i = \begin{cases} 1 & \text{if } (x,y) \in S_1 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

Assume S_1 is given, we average $\sup_g |\hat{R_{S_1}}(f) - \hat{R_{S_2}}(f)|$ over **ALL** partitions of S in (S_1, S_2) we get the Rademacher complexity.

To visualize S_1 and S_2 as datasets, we can use the following visualization:



Intuition 2: Measures how well $\mathcal{F}$ can fit our noisy labels.

Rademacher Lemma:

$$\mathbb{E}_{S\sim P^N}\left[\text{UnRep}(\mathcal{F},S)\right] \;\leq\; 2\;\mathbb{E}_{S\sim P^N}\left[\hat{Rad}(\mathcal{G})\right]$$

Theorem (PAC with Rademacher):

Assume $|l(f(x),y)|\leq c$ for all (x,y) (for any loss, but must be bounded)

For all $f\in\mathcal{F}$, if $S\sim P^N$, with probability at least $1-\delta$

$$R(f)-\hat{R}_S(f) \;\leq\; 2\,\hat{Rad}_S(l\circ\mathcal{F})+4\,c\sqrt{\frac{2ln(\frac{4}{\delta})}{N}}$$

We conclude the PAC result:

$$R(f)-R(f_{\mathcal{F}}) \;\leq\; ? \qquad \text{(DO AS AN EXERCISE, unless scribe)}$$

Exercise: let $\mathcal{G} = \left\{z\longmapsto\alpha : \alpha\in[-1,1]\right\}$

(1) What is $\hat{Rad}_S(\mathcal{G}) = ?$ (empirical Rademacher)

$$\sup_{\alpha\,\in\,[-1,1]}\sum_{i=1}^{N}\sigma_i\,\alpha \;=\; \sup_{\alpha\,\in\,\{-1,1\}}\sum_{i=1}^{N}\sigma_i\,\alpha \;=\; \left|\sum_{i=1}^{N}\sigma_i\right|$$

(α is the same for every σ_i)

$$\begin{aligned}
\Rightarrow\quad \hat{Rad}_S(G) &= \frac{1}{N}\,\mathbb{E}_{\sigma_1\ldots\sigma_N}\left|\sum_{i=1}^{N}\sigma_i\right| \\
&= \frac{1}{N}\,\mathbb{E}_{\sigma_i\ldots\sigma_N}\sqrt{\left(\sum\sigma_1\right)^2} \\
&\leq \frac{1}{N}\sqrt{\mathbb{E}_{\sigma_i\ldots\sigma_N}\left(\sum\sigma_1\right)^2} \\
&\quad\text{(Jensen Inequality *)} \\
&= \frac{1}{N}\sqrt{(\text{Var}(\sum\sigma_i)} \\
&= \frac{1}{N}\sqrt{N\times\text{Var}(\sigma_i)} \\
&= \frac{1}{\sqrt{N}}
\end{aligned}$$

* Jensen's Inequality states that if a function g is concave, then $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$. Here, the function g is the square root, which is a concave function. Thus Jensen's inequality can be applied and seen in:

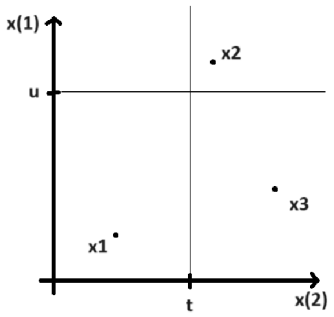
$$\frac{1}{N}\ \mathbb{E}_{\sigma_1 \ldots \sigma_N} \sqrt{\left(\sum \sigma_i\right)^2} \ \leq \ \frac{1}{N} \sqrt{\mathbb{E}_{\sigma_i \ldots \sigma_N} \left(\sum \sigma_1\right)^2}$$

Note: The σ's here are independently and identically distributed as stated by the definition of Rademacher, which allows us to linearize the variance in:

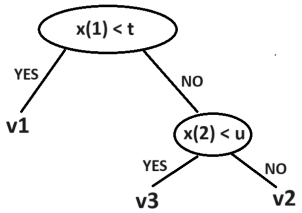
$$\frac{1}{N}\sqrt{(\mathrm{Var}(\sum \sigma_i)} \ = \ \frac{1}{N}\sqrt{N \times \mathrm{Var}(\sigma_i)}$$

(2) Let $\mathcal{G} \ = \ \left\{ \text{ decision trees which can output 1 or -1 at the leaves } \right\}$
What is $\hat{Rad}_S(\mathcal{G}) \ = \ ?$

Let us choose a dataset of 3 points.



(a) 3 random points



(b) Decision tree

Call T(j, k, l) this tree where $v_1 = j, v_2 = k, v_3 = l$.

σ_1	σ_2	σ_3	$\mathcal{G}$	$\sum \sigma_i\, g(x_i)$
-1	-1	-1	T(-1, -1, -1)	-3
-1	-1	1	T(-1, -1, 1)	-1
-1	1	-1	T(-1, 1, -1)	-1
-1	1	1	T(-1, 1, 1)	1
1	-1	-1	T(1, -1, -1)	-1
1	-1	1	T(1, -1, 1)	1
1	1	-1	T(1, 1, -1)	1
1	1	1	T(1, 1, 1)	3

Table 1: **Evaluation of Decision Tree Outputs**

Here, $\forall\, \sigma_1, ..., \sigma_N, \ \ \sup_{g \in \mathcal{G}} \sum_{i=1}^{N} \sigma_i\, g(x_i) = N$

This results in:

$$\boxed{\hat{Rad}_S(\mathcal{G}) \ = \ 1}$$

Which is BAD ! Thus, in the previous **PAC with Rademacher Theorem**, this bound is completely useless.

Theorem: For any $\mathcal{P}$, for $\mathcal{F}$ = linear models : $\mathcal{F} = \{x \longmapsto x^T\theta;\ \|\theta\|_2 \leq w_2\}$, and for any L-Lipschitz loss (hinge, logistic)

$$\text{UnRep}(\mathcal{F}, S) \ \leq \ \frac{w_2 x_2}{\sqrt{N}} \ + \ 4\, x_2 \sqrt{\frac{2ln(\frac{2}{\delta})}{N}}$$

$$\text{where } \frac{w_2 x_2}{\sqrt{N}} \text{ is the Rademacher complexity,}$$

$$\text{and } x_2 = \sup_{x \in X} \|x\|_2.$$

Now, the bound is not useless !

5 Conclusion

In this lecture we recapped the Hoeffding Inequality, as well as saw other formulas to express the inequality. We learned that the Hoeffding Ineqaulity gives a bound on the probability that the empirical error (error on the training data) deviates from the true error (error on the entire distribution) by at least t. This is used in PAC (Probably Approximately Correct) learning theory to provide guarantees on the generalization ability of algorithms.
We also learned about the Rademacher Complexity, which captures the expressiveness of the class $\mathcal{F}$. We saw that a higher Rademacher Complexity indicates a more expressive class that can potentially overfit to data, while a lower complexity suggests a less expressive class that might be more resistant to overfitting. We also explored the case of the unrepresentativeness of of $\mathcal{F}$ and S when $|\mathcal{F}| = \infty$.
Along the way, we used various examples and exercises to demonstrate all of this.