FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

Introduction to Statistical Learning Theory

Dr. Leonardo Rey Vega

University of Buenos Aires

October 2018

Outline

Learning model I

We will introduce a formal mathematical model for learning. We will need the following elements:

- Feature set $\mathcal{X}$: This set contains the elements that are observable at testing time and processed by our classifier or predictor, e.g., in classification tasks they the objects we want to label. The also receive the name of *examples*.
- Label set $\mathcal{Y}$: This set contains the labels in the case of a classification task or the variables we want to predict in the case of regression problems.
- Training set $\mathcal{S}^n$: $\mathcal{S}^n \equiv \{(x_i, y_i)\}_{i=1}^n$ is the training sequence that will use our algorithm or *learner* to generate our final classifier or predictor.
- The learning algorithm $\mathcal{T}(\mathcal{S}^n)$: This the device that takes the training set $\mathcal{S}^n$ generate the final classifier or predictor. Technically it is an operator $\mathcal{T} : \mathcal{S}^n \to \mathcal{H}$ where $\mathcal{H}$ is the *hypothesis space* where the prototypes of classifier or predictor live, that is every $h \in \mathcal{H}$ defines a function $h : \mathcal{X} \to \mathcal{Y}$.
- Data-generation model P_{XY}: We assume that training and testing samples are generated in i.i.d. fashion according to an arbitrary probability law P_{XY} totally unknown at training time.
- *Loss* or *risk* function ℓ: The loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is an appropriate function which for every hypothesis evaluate the loss that results from use h on $x \in \mathcal{X}$ to predict $y \in \mathcal{Y}$. Mathematically, this is can also be written as $\ell(h, (x, y)) \equiv \ell(h(x), y)$.

Learning model II

Some important loss functions:

- $\{0,1\}-$loss: In this case $\mathcal{Y} = \{0,1\}$, $\mathcal{X}$ is arbitrary, and $\mathcal{H}$ is set of functions $h : \mathcal{X} \to \{0,1\}$. The loss is defined as:

$$\ell(h,(x,y)) \equiv \mathbb{1}\{y \neq h(x)\}$$

- Square loss: In this case $\mathcal{Y}$ should be a normed linear space and $\mathcal{X}$ is arbitrary. The loss is defined as:

$$\ell(h,(x,y)) = \|y - h(x)\|^2$$

Assuming that the training set is i.i.d. according to some probability law P_{XY} is it is clear that $\hat{h} \equiv \mathcal{T}(\mathcal{S}^n)$ is a random variable that depends on the realization of $\mathcal{S}^n$. According to our loss function we would be interested in the *goodness* of this solution to the learning problem:

$$\mathcal{L}(\mathcal{T}(\mathcal{S}^n)) \equiv \mathbb{E}_{XY}\left[\ell(\hat{h}(X),Y)\right] = \int_{X\times\mathcal{Y}} \ell(\hat{h}(x),y)dP_{XY}$$

Again this is random variable depending on the realization of $\mathcal{S}^n$ and it is known as the *generalization*.

Learning model III

What about the algorithm $\mathcal{T}$? In this lecture we will analyze the *empirical risk minimization criterion* (ERM). That is:

$$\mathcal{T}\left(\mathcal{S}^n\right) = \hat{h}(\mathcal{S}^n) = \arg\min_{h\in\mathcal{H}} \hat{\mathcal{L}}(h), \tag{1}$$

with $\hat{\mathcal{L}}(h) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$ the *empirical risk*. Some comments:

- Other possible criterion can be cast as:

$$\hat{h}(\mathcal{S}^n) = \arg\min_{h\in\mathcal{H}} \hat{\mathcal{L}}(h) + f(h)$$

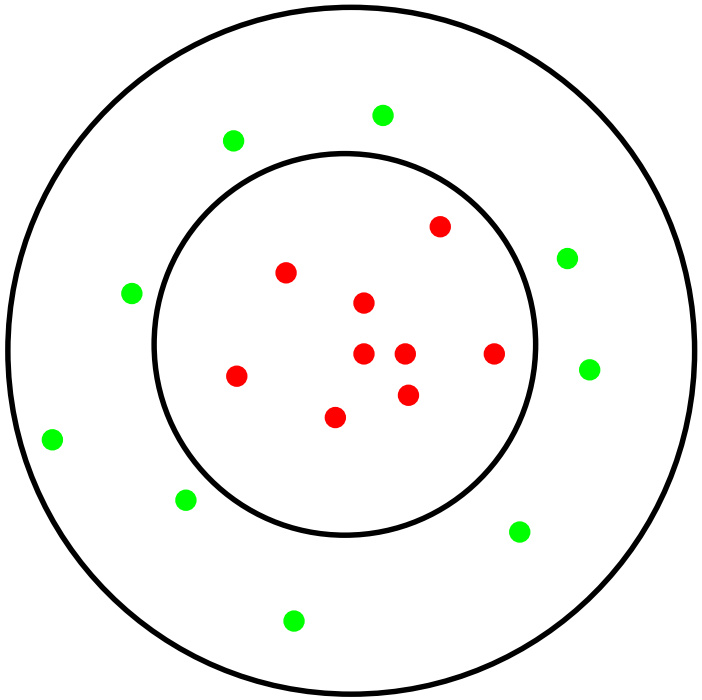
  where $f : \mathcal{H} \to \mathbb{R}_{\geq 0}$ is functional that includes a penalization by selecting a particular member of family $\mathcal{H}$. This is the concept of *regularization*.
- Solving (1) could be extremely difficult depending on ℓ and $\mathcal{H}$. In general some approximate solution can be expected but not the optimal one. We will not consider this issue and assume that we have the computational power to solve (1) exactly.

ERM could fail! I

Supose we have a model in which the examples X are distributed uniformly in circle of unit radius. Given X the labels (0 or 1) are determined by:

$$Y = \begin{cases} 1 & \|X\| \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Assume we have a set of examples with its corresponding labels as shown.



ERM could fail! II

Assume the $\{0,1\}-$ loss a consider ERM over all possible functions from the disk $\mathcal{B}(0,1)$ to $\{0,1\}$:

$$\hat{h} = \arg \min_{h:\mathcal{B}(0,1)\to\{0,1\}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \neq h(\mathbf{x}_i)\}$$

The following is a minimizer of the empirical loss:

$$\hat{h}(\mathbf{x}) = \begin{cases} y_i & \text{if exists such that } \mathbf{x} = \mathbf{x}_i \\ 0 & \text{otherwise} \end{cases}$$

- Note that with this choice $\mathcal{L}(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{y_i \neq \hat{h}(\mathbf{x}_i)\right\} = 0$.
- However

$$\begin{aligned}
\mathcal{L}(\hat{h}) &= \frac{1}{\pi}\int_{\mathcal{B}(0,1/2)} \mathbb{1}\left\{\hat{h}(\mathbf{x}) \neq 1\right\} d\mathbf{x} + \frac{1}{\pi}\int_{\mathcal{B}(0,1)\cap\mathcal{B}^c(0,1/2)} \mathbb{1}\left\{\hat{h}(\mathbf{x}) \neq 0\right\} d\mathbf{x} \\
&= \frac{1}{\pi}\int_{\mathcal{B}(0,1/2)} d\mathbf{x} \\
&= \frac{1}{4} \longrightarrow \text{Poor performance for every } n\text{!!!}
\end{aligned}$$

ERM could fail! III

Some comments:

- We have found a predictor which is optimal in terms of ERM but have an extremely poor average performance on new samples.
- This is the problem of *overfitting*. We have a predictor which *fits too well* the training samples ($\mathcal{L}(\hat{h}) = 0$). In fact, it memorizes the training samples.
- The problem is the hypothesis space $\mathcal{H}$ is *too big* (every function from the unit disk to $\{0, 1\}$).
- Even if we restrict our hypothesis space $\mathcal{H}$ to some *smaller* sets of functions we can have the same problem. For example consider $\mathcal{H}$ to be the set of functions such that:

$$h(\mathbf{x}) = \begin{cases} 1 & p(\mathbf{x}) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and where $p(\mathbf{x})$ is a polynomial. For every training set, we can find a polynomial $\hat{p}(\mathbf{x})$ such that the corresponding $\hat{h}(\mathbf{x})$ has $\hat{\mathcal{L}}(\hat{h}) = 0,\ \mathcal{L}(\hat{h}) = \frac{1}{4}$.

We conclude that the choice of $\mathcal{H}$ should be done carefully and a major question in learning theory is to identify which classes of functions $\mathcal{H}$ will not result in overfitting using the ERM criterion.

Bias-Variance trade-off I

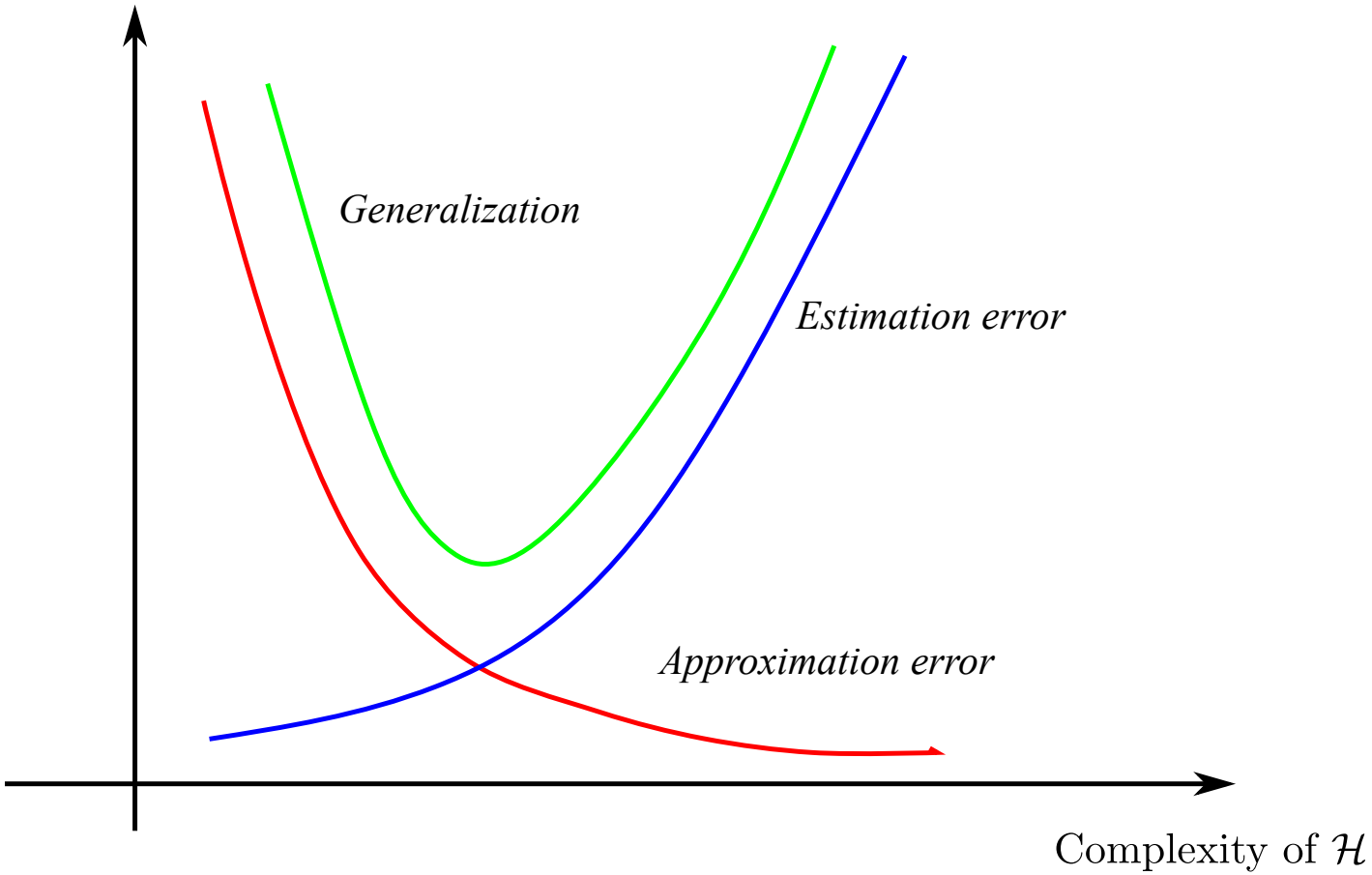
The generalization can be written in the following manner:

$$\mathcal{L}(\hat{h}) = \underbrace{\mathcal{L}(\hat{h}) - \min_{h \in \mathcal{H}} \mathcal{L}(h)}_{\mathcal{E}_{\text{est}}(\hat{h}, \mathcal{H})} + \overbrace{\min_{h \in \mathcal{H}} \mathcal{L}(h)}^{\mathcal{E}_{\text{approx}}(\mathcal{H})}$$

- The *approximation* error $\mathcal{E}_{\text{approx}}(\mathcal{H})$ depends *only* on the class of hypothesis $\mathcal{H}$ and not on the training sample $\mathcal{S}^n$. It is expected that if $\mathcal{H}_1 \subseteq \mathcal{H}_2$ then $\mathcal{E}_{\text{approx}}(\mathcal{H}_2) \leq \mathcal{E}_{\text{approx}}(\mathcal{H}_1)$. Clearly $\mathcal{E}_{\text{approx}}(\mathcal{H}) \geq \mathcal{L}^*$ where $\mathcal{L}^*$ is the minimal value of risk when $\mathcal{H}$ is composed by all possible functions (Bayes error for classification and minimum square error for mean-square error regression).
- The *estimation* error $\mathcal{E}_{\text{est}}(\hat{h}, \mathcal{H})$ depends on both, the hypothesis family and the training sample. Its behaviour is more complex, however it is not difficult to see that for fixed training sample $\mathcal{S}^n$, if if $\mathcal{H}_1 \subseteq \mathcal{H}_2$ then $\mathcal{E}_{\text{est}}(\hat{h}, \mathcal{H}_2) \geq \mathcal{E}_{\text{est}}(\hat{h}, \mathcal{H}_1)$

This leads us to an interesting trade-off which is known as the *bias-variance* trade-off.

Bias-Variance trade-off II



PAC criterion I

The generalization $\mathcal{L}(\hat{h})$ (or the estimation error $\mathcal{E}_{\text{est}}(\hat{h}, \mathcal{H})$) depends on the realization of $\mathcal{S}^n$ in i.i.d. fashion according to some unknown probability distribution P_{XY}.

In this sense the generalization is a random variable, whose ultimate value would depends (in a rather complicated manner) on the actual $\mathcal{S}^n$ sampled.

As we could expect that some values of $\mathcal{S}^n$ be extremely bad for learning a given task we want to have some probabilistic *guarantees* for sucessful learning.

And we want to achieve this being agnostic on the true distribution of data P_{XY}!!

Definition (PAC Learnability of hypothesis family $\mathcal{H}$)

We said that family $\mathcal{H}$ is PAC-learnable if for every $\epsilon, \delta > 0$ exists a learning algorithm $\mathcal{T}$ and function $N(\epsilon, \delta) : (0,1) \times (0,1) \to \mathbb{N}$ such that for every distribution P_{XY} if the number of training samples satisfies $n \geq N(\epsilon, \delta)$

$$\mathbb{P}\left(\mathcal{S}^n : \mathcal{L}(\hat{h}) - \min_{h \in \mathcal{H}} \mathcal{L}(h) \leq \epsilon\right) \geq 1 - \delta$$

PAC criterion II

Some comments:

- The definition includes two accuracy parameters. Parameter ϵ is related to the possibility that a given learning algorithm, with a training sequence $\mathcal{S}^n$, have of achieving a generalization error close to the best in the hypothesis class (approximately correct part!!). Parameter δ quantifies the probability that we get a nice training sequence for accomplish this (probably part!!).
- A PAC learnable algorithm with hypothesis class $\mathcal{H}$ has universal (for all possible data probability distribution P_{XY}) performance guarantees as long as the number of training samples satisfies $n \geq N(\epsilon, \delta)$.
- For this reason $N(\epsilon, \delta)$ is known as the *sample complexity* of the algorithm.
- We can check with this definition if any possible learning algorithm could *learn* a task using the hypothesis family $\mathcal{H}$.
- We would only analyze the ERM algorithm and what hypothesis classes could be used efficiently with this criterion for truly learning from real-world examples.

The case of finite $\mathcal{H}$ I

Lets begin with a very simple (although not a very realistic) scenario:

- $\ell(y', y) \leq \alpha$ for all $y', y \in \mathcal{Y}$ and $\ell(y, y) = 0$ for all $y \in \mathcal{Y}$.
- $\mathcal{H}$ is finite, that is $|\mathcal{H}| < \infty$.
- Restrict the possible data generating distribution to the following family. Let us consider any possible data distribution P_X for the examples x. However we will assume that exists (although totally unknown to us) $f^* : \mathcal{X} \times \mathcal{Y}$ such that for every P_X we have

$$Y = f^*(X)$$

  That is, there is no *noise* in Y given X. Any variability in the training sample is due only to the examples in $\mathcal{X}$.
- Finally we are extremely luckily in the sense that $f^* \in \mathcal{H}$.

Not difficult to see that (for $\{0, 1\}-$loss or the square loss):

$$\mathcal{L}(f^*) = 0$$

and obviously $\mathcal{E}_{\text{approx}}(\mathcal{H}) = \min_{h \in \mathcal{H}} \mathcal{L}(h) = 0$.

The case of finite $\mathcal{H}$ II

For the ERM algorithm we want to find $N(\epsilon, \delta)$ such that for $n \geq N(\epsilon, \delta)$:

$$\mathbb{P}\left\{\mathcal{S}^n : \mathcal{L}(\hat{h}) \leq \epsilon\right\} \geq 1 - \delta$$

It is clear that the $\hat{h}$ given by ERM algorithm must satisfy $\hat{\mathcal{L}}(\hat{h}) = 0$.

The main issue here it that the solution to the minimization to the ERM could not be the true f^* and for that reason $\mathcal{L}\left(\hat{h}\right) \neq 0$.

Define

$$\mathcal{H}_F \equiv \{h \in \mathcal{H} : \mathcal{L}(h) > \epsilon\}$$

$$\mathcal{A} \equiv \left\{\mathcal{S}^n : \exists h \in \mathcal{H}_B,\ \hat{\mathcal{L}}(h) = 0\right\}$$

We say that $\mathcal{H}_F$ is the set of $\epsilon-$bad hypothesis and $\mathcal{A}$ is the set of ϵ-bad training sets.

Clearly:

$$\mathbb{P}\left\{\mathcal{S}^n : \mathcal{L}(\hat{h}) > \epsilon\right\} = \mathbb{P}\left\{\mathcal{S}^n : \hat{h} = h \text{ for some } h \in \mathcal{H}_F\right\} \leq \mathbb{P}\{\mathcal{A}\}$$

The case of finite $\mathcal{H}$ III

Using the union bound ($\mathbb{P}\left(\cup_i \mathcal{B}_i\right) \leq \sum_i \mathbb{P}(\mathcal{B}_i)$) we can write:

$$\mathbb{P}\left\{\mathcal{A}\right\} = \mathbb{P}\left\{\bigcup_{h\in\mathcal{H}_F}\left\{\mathcal{S}^n : \hat{\mathcal{L}}(h) = 0\right\}\right\} \leq \sum_{h\in\mathcal{H}_F} \mathbb{P}\left\{\mathcal{S}^n : \hat{\mathcal{L}}(h) = 0\right\}$$

However:

$$\hat{\mathcal{L}}(h) = 0 \Leftrightarrow h(x_i) = f^*(x_i)\ \forall i = 1, \ldots, n$$

and for $h \in \mathbb{H}_F$

$$\mathbb{P}\left\{\mathcal{S}^n : \hat{\mathcal{L}}(h) = 0\right\} = \prod_{i=1}^{n} \mathbb{P}\left\{x_i : h(x_i) = f^*(x_i)\right\} = \left(\mathbb{P}\left\{x : h(x) = f^*(x)\right\}\right)^n$$

Clearly

$$\begin{aligned}\mathcal{L}(h) &= \mathbb{E}\left[\ell(h(X), f^*(X))\right] \\ &= \int_{\{x:h(x)\neq f^*(x)\}} \ell(h(x), f^*(x)) dP_X \\ &\leq \alpha \mathbb{P}\left\{x : h(x) \neq f(x)\right\}\end{aligned}$$

The case of finite $\mathcal{H}$ IV

As $h \in \mathcal{H}_F$:

$$\mathbb{P}\left\{x : h(x) = f^*(x)\right\} \leq 1 - \frac{\mathcal{L}(h)}{\alpha} \leq 1 - \frac{\epsilon}{\alpha}$$

we can write:

$$\mathbb{P}\left\{\mathcal{S}^n : \hat{\mathcal{L}}(h) = 0\right\} \leq \left(1 - \frac{\epsilon}{\alpha}\right)^n \leq e^{-n\frac{\epsilon}{\alpha}}$$

Combining the results we get:

$$\mathbb{P}\left\{\mathcal{S}^n : \mathcal{L}(\hat{h}) > \epsilon\right\} \leq |\mathcal{H}_F| e^{-n\frac{\epsilon}{\alpha}} \leq |\mathcal{H}| e^{-n\frac{\epsilon}{\alpha}}$$

or what it is the same:

$$\mathbb{P}\left\{\mathcal{S}^n : \mathcal{L}(\hat{h}) \leq \epsilon\right\} \geq 1 - |\mathcal{H}| e^{-n\frac{\epsilon}{\alpha}}$$

If $\delta \equiv |\mathcal{H}| e^{-n\frac{\epsilon}{\alpha}}$ we get the following sample complexity:

$$N(\epsilon, \delta) = \frac{\alpha}{\epsilon} \log\left(\frac{|\mathcal{H}|}{\delta}\right)$$

The case of finite $\mathcal{H}$ V

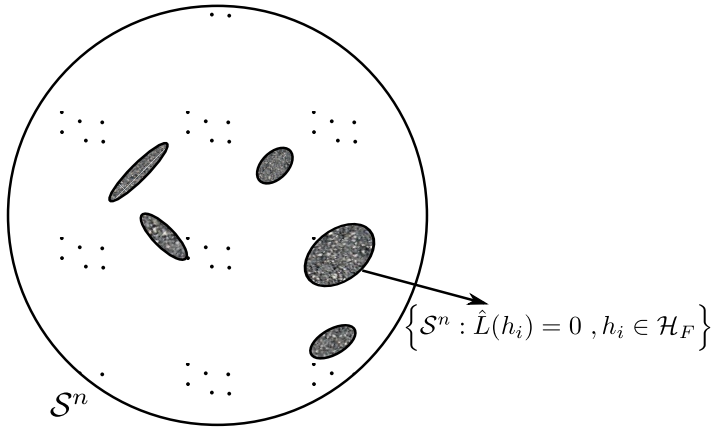
We have proved the following theorem:

Theorem (PAC learnability for finite $\mathcal{H}$ (weak version))

Consider a finite hypothesis class $\mathcal{H}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ which is bounded by α and satisfies $\ell(y, y) = 0$ for all $y \in \mathcal{Y}$. For every distribution on $\mathcal{X}$ such that $Y = f^*(X)$ with $f^* \in \mathcal{H}$ we obtain

$$\mathbb{P}\left\{\mathcal{S}^n : \mathcal{L}(\hat{h}) \leq \epsilon\right\} \geq 1 - \delta$$

for $n \geq \frac{\alpha}{\epsilon} \log\left(\frac{|\mathcal{H}|}{\delta}\right)$.



Application of union bound to family $\mathcal{H}_F$.

A brief detour: concentration inequalities I

In order to continue with our exposition we need to introduce some results on functions of independent random variables.
We are all familiarized with the following: If $\{X_i\}_{i=1}^{\infty}$ is sequence of independent and identically distributed random variables with $\mathbb{E}[|X|] < \infty$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}[X] \text{ with probability 1}$$

- Asymptotic result with minimal requirements.
- Result valid for a particular *linear function* of the sequence of random variables. What about more general functions??
- Moreover, we do not have information about the *tails* of the converging distribution, that is information about

$$\mathbb{P}\left\{\left|f(X_1, X_2, \ldots, X_n) - \mathbb{E}\left[f(X_1, X_2, \ldots, X_n)\right]\right| > \epsilon\right\}$$

  as function of n and ϵ.

These questions can be answered by results in the area of concentration of measures inequalities.

A brief detour: concentration inequalities II

Some well-known tails bounds:

- Markov inequality: If X is a non-negative random variable:

$$\mathbb{P}\{X \geq \epsilon\} \leq \frac{\mathbb{E}[X]}{\epsilon}$$

- Tschebyschev inequality: For any random variable X with finite variance:

$$\mathbb{P}\{|X - E[X]| \geq \epsilon\} \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$$

- Chernoff bound: If X is random variable such that exists $s > 0$ such that $\mathbb{E}\left[e^{sX}\right] < \infty$

$$\mathbb{P}\{X \geq \epsilon\} \leq \inf_{s>0} e^{-s\epsilon}\mathbb{E}\left[e^{sX}\right]$$

  This bound is extremely powerful and give very tight results. However, we need information about $\mathbb{E}\left[e^{sX}\right]$ (e.g. sub-gaussianity!!) and be able to solve for the infimum.

- More sophisticated bounds from *isoperimetric inequalities*, *entropy inequalities*, *transport inequalities*, etc.

A brief detour: concentration inequalities III

We present the following result (without proof):

Theorem (McDiarmid (1989))

Consider a sequence of independent random variables $\{X_i\}_{i=1}^{\infty}$ *taking values in an arbitrary set* $\mathcal{X}$*. Consider a function* $f : \mathcal{X}^n \to \mathbb{R}$ *that satisfies:*

$$\sup_{(x_1,x_2,\ldots,x_i,x_i',\ldots x_n)\in\mathcal{X}^{n+1}} |f(x_1,x_2,\ldots,x_i,\ldots,x_n)-f(x_1,x_2,\ldots,x_i',\ldots,x_n)| \leq \alpha_i$$

for $i = 1,\ldots,n$*. Then:*

$$\mathbb{P}\left\{f(X_1,\ldots X_n) - \mathbb{E}\left[f(X_1,\ldots,X_n)\right] > \epsilon\right\} \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n \alpha_i^2}\right)$$

$$\mathbb{P}\left\{f(X_1,\ldots X_n) - \mathbb{E}\left[f(X_1,\ldots,X_n)\right] < -\epsilon\right\} \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n \alpha_i^2}\right)$$

A brief detour: concentration inequalities IV

McDiarmid inequality is sometimes known as the *bounded differences* inequality.

The basic principle behind McDiarmid result and almost all concentration inequalities is that "a random variable that smoothly depends on the influence of many independent random variables satisfy exponential bounds for its tail probabilities" (Talagrand, 1995).

A very well-known special case of the McDiarmid inequality is the following:

Theorem (Hoeffding (1963))

Consider $X_1, \ldots, X_n$ *independent bounded random variables such that* $X_i \in [a_i, b_i]$. *For any* $\epsilon > 0$*:*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| > \epsilon\right\} \leq 2\exp\left(-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Finite $\mathcal{H}$ again I

We are going to remove the assumption about the existence of f^* such that $f^*(X) = Y$. In this manner, we will consider completely general distributions P_{XY}. We will continue, however with the finite assumption for $\mathcal{H}$ and all the previous assumptions.

Obviously, in this case we may have

$$\mathcal{E}_{\text{approx}}(\mathcal{H}) = \min_{h \in \mathcal{H}} \mathcal{L}(h) \neq 0$$

and

$$\hat{\mathcal{L}}(\hat{h}) \neq 0$$

Our arguments to show PAC learnability of $\mathcal{H}$ under this setting should have slightly different with respect to the later case. Remember that we need to find $N(\epsilon, \delta) : (0,1) \times (0,1) \to \mathbb{N}$ such that for every distribution P_{XY} if the number of training samples satisfies $n \geq N(\epsilon, \delta)$

$$\mathbb{P}\left(\mathcal{S}^n : \mathcal{L}(\hat{h}) - \min_{h \in \mathcal{H}} \mathcal{L}(h) \leq \epsilon\right) \geq 1 - \delta$$

Finite $\mathcal{H}$ again II

Lemma (Uniform bounding of $\mathcal{L}(\hat{h}) - \min_{h\in\mathcal{H}} \mathcal{L}(h)$)

For any hypothesis family $\mathcal{H}$ the following is true:

$$\mathcal{L}(\hat{h}) - \min_{h\in\mathcal{H}} \mathcal{L}(h) \leq 2 \sup_{h\in\mathcal{H}} \left|\hat{\mathcal{L}}(h) - \mathcal{L}(h)\right|$$

Proof: We can write:

$$\begin{aligned}
\mathcal{L}(\hat{h}) - \min_{h\in\mathcal{H}} \mathcal{L}(h) &= \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) - \min_{h\in\mathcal{H}} \mathcal{L}(h) + \hat{\mathcal{L}}(\hat{h}) \\
&\leq \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) + \sup_{h\in\mathcal{H}} \left|\hat{\mathcal{L}}(h) - \mathcal{L}(h)\right| \\
&\leq 2 \sup_{h\in\mathcal{H}} \left|\hat{\mathcal{L}}(h) - \mathcal{L}(h)\right|
\end{aligned}$$

This lemma allow us to write:

$$\mathbb{P}\left(\mathcal{S}^n : \mathcal{L}(\hat{h}) - \min_{h\in\mathcal{H}} \mathcal{L}(h) > \epsilon\right) \leq \mathbb{P}\left(\mathcal{S}^n : 2 \sup_{h\in\mathcal{H}} \left|\hat{\mathcal{L}}(h) - \mathcal{L}(h)\right| > \epsilon\right)$$

Finite $\mathcal{H}$ again III

As the family $\mathcal{H}$ is finite we can write:

$$\mathbb{P}\left(\mathcal{S}^n : 2\sup_{h\in\mathcal{H}}\left|\hat{\mathcal{L}}(h)-\mathcal{L}(h)\right|>\epsilon\right)\leq\sum_{h\in\mathcal{H}}\mathbb{P}\left\{\mathcal{S}^n : \left|\hat{\mathcal{L}}(h)-\mathcal{L}(h)\right|>\frac{\epsilon}{2}\right\}$$

Notice that for every $h\in\mathcal{H}$ we have that:

$$\hat{\mathcal{L}}(h)=\frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i);y_i),\ \mathbb{E}\left[\hat{\mathcal{L}}(h)\right]=\mathcal{L}(h)$$

As, the loss function $\ell(\cdot,\cdot)$ is bounded by α we can use Hoeffding inequality to show that

$$\mathbb{P}\left\{\mathcal{S}^n : \left|\hat{\mathcal{L}}(h)-\mathcal{L}(h)\right|>\frac{\epsilon}{2}\right\}\leq 2\exp\left(-\frac{\epsilon^2 n}{2\alpha^2}\right)$$

and

$$\mathbb{P}\left(\mathcal{S}^n : 2\sup_{h\in\mathcal{H}}\left|\hat{\mathcal{L}}(h)-\mathcal{L}(h)\right|>\epsilon\right)\leq 2|\mathcal{H}|\exp\left(-\frac{\epsilon^2 n}{2\alpha^2}\right)$$

Finite $\mathcal{H}$ again IV

Theorem (PAC learnability for finite $\mathcal{H}$ (strong version))

Consider a finite hypothesis class $\mathcal{H}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ which is bounded by α and satisfies $\ell(y, y) = 0$ for all $y \in \mathcal{Y}$. For every distribution on $\mathcal{X} \times \mathcal{Y}$

$$\mathbb{P}\left\{\mathcal{S}^n : \mathcal{L}(\hat{h}) - \min_{h \in \mathcal{H}} \mathcal{L}(h) \leq \epsilon\right\} \geq 1 - \delta$$

for $n \geq \frac{2\alpha^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$.

- It is important the emphasize that the result is valid for *every possible* distribution of data.
- Compare the sample complexity in this case with the corresponding the weak version which assume the existence of f^* given by:

$$N(\epsilon, \delta) = \frac{\alpha}{\epsilon} \log\left(\frac{|\mathcal{H}|}{\delta}\right)$$

No free lunch theorem

As we saw finite family of hypothesis are PAC learnable. What about infinite families?

This is a very delicate question that will briefly analyze in a while. But we can be ambitious and consider the learnability of the complete set of functions from $\mathcal{X} \times \mathcal{Y}$:

$$\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$$

It is clear that $|\mathcal{H}| = |\mathcal{Y}|^{|\mathcal{X}|}$ which show us that if $\mathcal{X}$ and/or $\mathcal{Y}$ is infinite then $\mathcal{H}$ is infinite. Unfortunately we have the following result (presented without proof):

Theorem (No free-lunch theorem for learning)

The class $\mathcal{H}$ of all functions $h : \mathcal{X} \to \mathcal{Y}$ is not PAC learnable for infinite $\mathcal{X}$ and/or $\mathcal{Y}$ for bounded loss functions $\ell : \mathcal{Y} \to \mathcal{Y}$.

The theorem basically says that there is not *finite* sample complexity for successful learning over the set all possible distributions if we have not inductive bias for choosing the hypothesis set $\mathcal{H}$.

Learning infinite $\mathcal{H}$ I

In the following for simplicity we will restrict to the case with the $\{0,1\}$-loss.

Although the class of all functions from $\mathcal{X}$ to $\mathcal{Y}$ is not learnable, we will show that if the *complexity* of the class $\mathcal{H}$ is not *large* successful learning can be achieved even if $\mathcal{H}$ is infinite.

This will show that the size of class $\mathcal{H}$ is not the key factor in the its learnability.

Let us consider a very simple example. Assume that we are considering $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0,1\}$ with the $\{0,1\}$- loss. We will consider the set of joint distributions for (X,Y) given by:

$$\mathcal{P} = \{P_{XY} : P_X(x) \text{ is arbitrary, } Y = \mathbb{1}\{X < \alpha^*\} \text{ where } \alpha^* \in \mathbb{R} \text{ is arbitrary}\}$$

We will consider the following class $\mathcal{H}$:

$$\mathcal{H} = \{\mathbb{1}\{x < a\} : a \in \mathbb{R}\}$$

It is clear that $\mathcal{E}_{\text{approx}} = 0$ and $\mathcal{H}$ is infinite.

Is this class learnable? The answer is yes!!

Learning infinite $\mathcal{H}$ II

Assume that we have a training data $\mathcal{S}^n = \{(x_i, y_i)\}_{i=1}^n$. We define:

$$\beta_1(\mathcal{S}^n) = \max\{x : (x, 1) \in \mathcal{S}^n\}$$

$$\beta_2(\mathcal{S}^n) = \min\{x : (x, 0) \in \mathcal{S}^n\}$$

It is clear that if work with the ERM criterion and choose $\hat{h} = \mathbb{1}\{x < \beta^*\}$ with $\beta^* \in (\beta_1(\mathcal{S}^n), \beta_2(\mathcal{S}^n))$ we obtain $\hat{\mathcal{L}}(\hat{h}) = 0$. For the training data distribution P_{XY} define α_1 and α_2 such that:

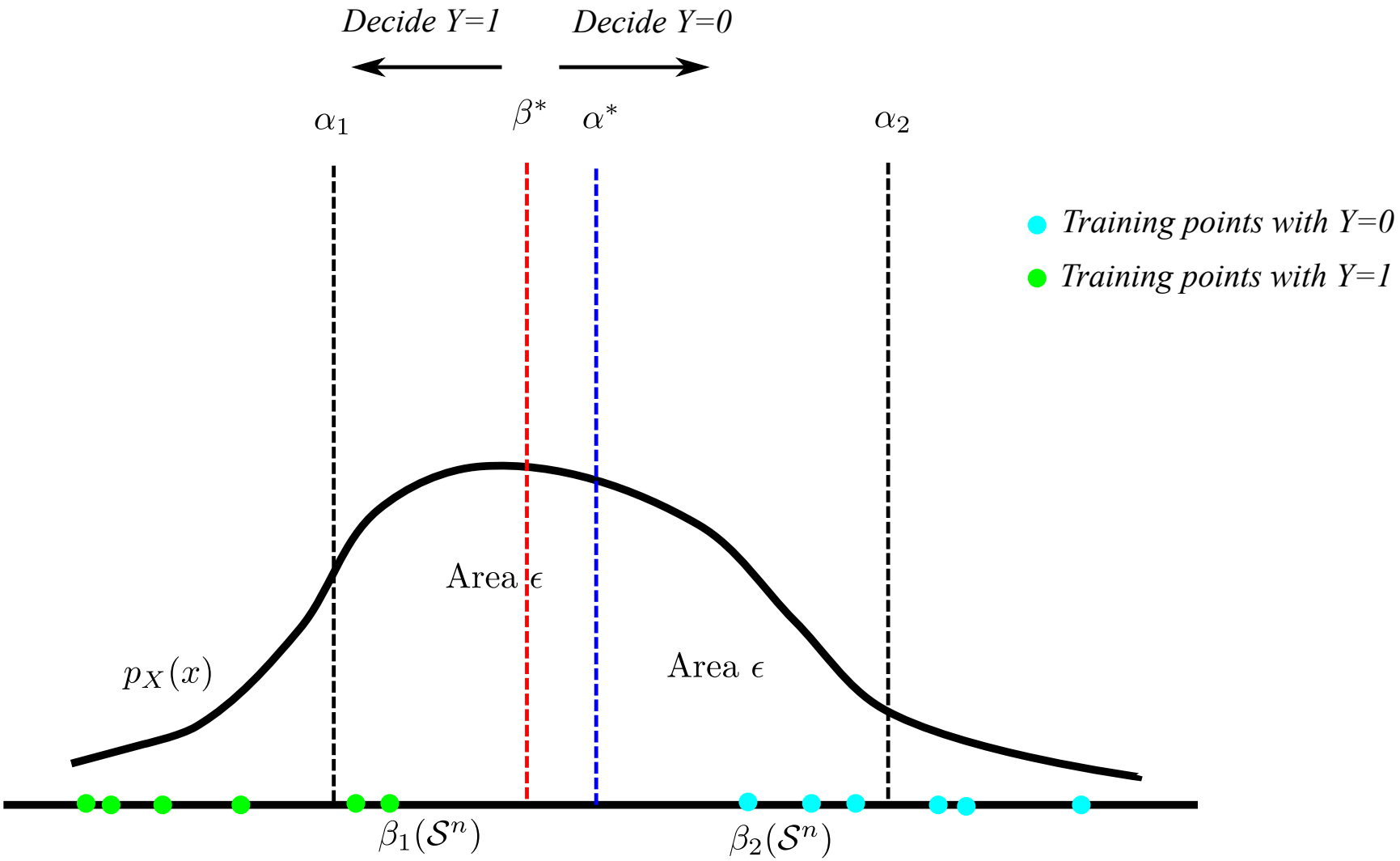
$$\mathbb{P}\{x \in (\alpha_1, \alpha^*)\} = \mathbb{P}\{x \in (\alpha^*, \alpha_2)\} = \epsilon$$

It is pretty clear that if $\beta_1(\mathcal{S}^n) \geq \alpha_1$ and $\beta_2(\mathcal{S}^n) \leq \alpha_2$ gives us that $\mathcal{L}(\hat{h}) \leq \epsilon$, which allow us to write:

$$\begin{aligned}\mathbb{P}\left\{\mathcal{S}^n : \mathcal{L}(\hat{h}) > \epsilon\right\} &\leq \mathbb{P}\{x_i \notin (\alpha_1, \alpha^*) : \forall i \in [1:n]\} + \mathbb{P}\{x_i \notin (\alpha^*, \alpha_2) : \forall i \in [1:n]\} \\ &\leq 2(1-\epsilon)^n \leq 2\exp(-n\epsilon)\end{aligned}$$

which gives $N(\epsilon, \delta) = \frac{\log(\frac{2}{\delta})}{\epsilon}$, which is finite sample complexity.

Learning infinite $\mathcal{H}$ III



VC dimension I

In previous example the key in family $\mathcal{H}$ is that it cannot discriminate between any pattern of different sample points.

- Assume that we have only one training point x_1. Then in family $\mathcal{H}$ we can choose $\beta \in \mathbb{R}$ such that $h = \mathbb{1}\{x < \beta\}$ discriminate perfectly the observed point:
  - Choose $\beta > x_1$ and $h(x_1) = 1$.
  - Choose $\beta < x_1$ and $h(x_1) = 0$.
- Assume that we have only two training points x_1 and x_2. Then in family $\mathcal{H}$ we cannot choose $\beta \in \mathbb{R}$ such that $h = \mathbb{1}\{x < \beta\}$ discriminate perfectly the observed points for some configuration of them. Assume that we sample x_1 and x_2 with $x_1 < x_2$:
  - If $\beta < x_1$, $h(x_1) = h(x_2) = 1$.
  - If $\beta > x_2$, $h(x_1) = h(x_2) = 0$.
  - If $x_1 < \beta < x_2$, $h(x_1) = 1$, $h(x_2) = 0$.
  - There is no choice of β such that $h(x_2) = 1$ and $h(x_1) = 0$.

With this set of hypothesis it is impossible to have maximal overfitting for every training points configuration when the training set satisfies $n \geq 2$.

VC dimension II

We can put the previous observation in mathematical terms:

Definition (Shattering)

We said that an hypothesis class *shatters* an arbitrary finite set $\mathcal{A} \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to $\mathcal{A}$ contains all the possible mappings from $\mathcal{A}$ to $\{0, 1\}$ (which are $2^{|\mathcal{A}|}$).

The critical parameter of class $\mathcal{H}$ for PAC learnability is the *Vapnik-Chervonenkis* (VC) dimension.

Definition (VC dimension of class $\mathcal{H}$, Vapnik-Chervonenkis (1970))

Given a hypothesis class $\mathcal{H}$ we define its VC dimension ($\text{VC}(\mathcal{H})$) as the maximal size of an arbitrary set $\mathcal{A} \in \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrary size we say that $\text{VC}(\mathcal{H}) = \infty$.

VC dimension III

Examples:

- A finite class $\mathcal{H}$ over an arbitrary space has $\mathrm{VC}(\mathcal{H}) \leq \log(|\mathcal{H}|)$.
- The class $\mathcal{H}$ of *threshold* function over $\mathbb{R}$, that is functions of the form $\mathbb{1}\{x < \beta\}$ with $\beta \in \mathbb{R}$ has $\mathrm{VC}(\mathcal{H}) = 1$.
- The class $\mathcal{H}$ of intervals over $\mathbb{R}$, that is functions of the form $\mathbb{1}\{\beta_1 < x < \beta_2\}$ with $\beta_1, \beta_2 \in \mathbb{R}$ has $\mathrm{VC}(\mathcal{H}) = 2$.
- The class $\mathcal{H}$ of rectangles in $\mathbb{R}^d$ with $d > 0$, that is functions $\mathbb{1}\{\beta_{1i} < x_i < \beta_{2i}\ i = 1, \ldots, d\}$ has $\mathrm{VC}(\mathcal{H}) = 2d$.
- The class of *half-spaces* in $\mathbb{R}^d$ with $d > 0$, that is functions $\mathbb{1}\left\{\mathbf{w}^T\mathbf{x} + b > 0\right\}$ has $\mathrm{VC}(\mathcal{H}) = d + 1$.
- The class of *convex polygons* in $\mathbb{R}^d$ with $d > 0$ has $\mathrm{VC}(\mathcal{H}) = \infty$.
- Consider a set of r linearly independent real functions $\{g_i(\mathbf{x})\}_{i=1}^r$ in $\mathbb{R}^d$ with $d > 0$. Then the set $\mathcal{H}$ of functions $\mathbb{1}\left\{\sum_{i=1}^r \alpha_i g_i(\mathbf{x}) > 0\right\}$ has $\mathrm{VC}(\mathcal{H}) \leq r$.
- The class $\mathcal{H}$ composed by neural networks with L layers, ReLU activation functions and W tunable parameters presents $\mathrm{VC}(\mathcal{H}) \leq \mathcal{O}(WL\log(W))$ (Bartlett *et al*, 2017).

PAC Learnability for general classes $\mathcal{H}$ I

Theorem (PAC Learnability for general classes $\mathcal{H}$)

Consider the $\{0,1\}$-loss with $\mathcal{X}$ arbitrary. The following are equivalent:

- *$\mathcal{H}$ is PAC learnable.*
- *ERM criterion is a PAC learner for $\mathcal{H}$.*
- *$VC(\mathcal{H}) < \infty$.*
- *The family $\mathcal{H}$ has the uniform convergence property: for every $\epsilon, \delta > 0$ exist $N^*(\epsilon, \delta)$ such that for all $n \geq N^*(\epsilon, \delta)$:*

$$\mathbb{P}\left\{\sup_{h \in \mathcal{H}} |\hat{\mathcal{L}}(h) - \mathcal{L}(h)| \leq \epsilon\right\} \geq 1 - \delta$$

Moreover, if $\mathcal{H}$ is PAC learnable the sample complexity $N(\epsilon, \delta)$ satisfies:

$$K_1 \frac{VC(\mathcal{H}) + \log(\frac{1}{\delta})}{\epsilon^2} \leq N(\epsilon, \delta) \leq K_2 \frac{VC(\mathcal{H}) + \log(\frac{1}{\delta})}{\epsilon^2}$$

for some positive constants K_1 and K_2.

PAC Learnability for general classes $\mathcal{H}$ II

Some comments:

- For the multi-class classification problem, the critical parameter is the *Natarajan* dimension (Natarajan, 1989) of family $\mathcal{H}$. The results for this case are pretty the same in the senses that Natarajan dimension is finite if and only if $\mathcal{H}$ is PAC learnable. However, in the multi-class classification problem uniform convergence property over $\mathcal{H}$ is not longer necessary.
- For regression problems with bounded square-loss and absolute-loss we have a similar result with an appropriate metric of complexity, similar to VC dimension, called the *fat shattering* dimension (Bartlett *et al*, 1994).
- Notice that en general, for the $\{0,1\}$-loss and for PAC learnable hypothesis families $\mathcal{H}$, we can write the following:

$$\mathcal{L}(h) \leq \hat{\mathcal{L}}(h) + \epsilon(\mathrm{VC}(\mathcal{H}), \delta, n) \text{ with probability } 1-\delta$$

for every $h \in \mathcal{H}$ with a function ϵ that depends on the class $\mathcal{H}$, δ and the number of samples n. This function quantifies the gap (as a function of n and the complexity of $\mathcal{H}$) between the empirical risk and the true one. In some sense, $\hat{\mathcal{L}}(h) + \epsilon(\mathrm{VC}(\mathcal{H}), \delta, n)$ can be interpreted as surrogate for the true risk.

Structural risk minimization

The previous comment will allows to introduce the idea of *Structural Risk Minimization* (SRM). Assume that our class $\mathcal{H}$ can be written as:

$$\mathcal{H} = \bigcup_{m \in \mathbb{N}} \mathcal{H}_m$$

where for each $m \in \mathbb{N}$, $\mathrm{VC}(\mathcal{H}_m) < \infty$ (notice that we could have $\mathrm{VC}(\mathcal{H}) = \infty$). Assuming that the we are working with the $\{0,1\}$-loss we have finite sample complexities $N_m(\epsilon, \delta_m)$ and $\epsilon_m(\mathrm{VC}(\mathcal{H}_m), \delta_m, n)$ for each $m \in \mathbb{N}$. Define also for every $h \in \mathcal{H}$

$$m(h) \equiv \min\{m : h \in \mathcal{H}\}$$

With the the size n of the training sample fixed we can set the following the following criterion for choosing our learner $\hat{h}$:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \hat{\mathcal{L}}(h) + \epsilon_{m(h)}(\mathrm{VC}(\mathcal{H}_{m(h)}), \delta_{m(h)}, n) \right\}$$

This is not an ERM problem. We are open to trading some loss of performance in the minimization of the empirical risk looking for a solution in class $\mathcal{H}_m$ which gives a lower value $\epsilon_m(\mathrm{VC}(\mathcal{H}_m), \delta_m, n)$. In some sense, we are proposing a regularization for the empirical risk minimization which takes into account the complexities of classes $\mathcal{H}_m$.

Bibliography

Basics of Statistical Learning Theory

- V. N. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.
- V. Vapnik, The Nature of Statistical Learning Theory, 2nd edition. New York: Springer, 1999.
- V. N. Vapnik, "An overview of statistical learning theory," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988-999, Sep. 1999.

More advanced texts on Statistical Learning Theory

- L. Devroye, L. Györfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Corrected edition. New York: Springer, 1997.
- S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, 1 edition. New York: Cambridge University Press, 2014.
- F. Cucker and S. Smale, "On the mathematical foundations of learning," Bull. Amer. Math. Soc., vol. 39, no. 1, pp. 1-49, 2002.
- D. Haussler, M. Kearns, and R. E. Schapire, "Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension," Machine Learning, vol. 14, no. 1, pp. 83-113, Jan. 1994.

Bibliography

More advanced texts on Statistical Learning Theory

- T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, "General conditions for predictivity in learning theory," Nature, vol. 428, Mar. 2004.
- D. A. McAllester, "PAC-Bayesian Model Averaging," in Proceedings of the Twelfth Annual Conference on Computational Learning Theory, New York, NY, USA, 1999, pp. 164-170.
- O. Bousquet and A. Elisseeff, "Stability and Generalization," J. Mach. Learn. Res., vol. 2, pp. 499-526, Mar. 2002.
- H. Xu and S. Mannor, "Robustness and generalization," Mach Learn, vol. 86, no. 3, pp. 391-423, Mar. 2012.
- R. M. Dudley, E. Giné, and J. Zinn, "Uniform and universal Glivenko-Cantelli classes," J Theor Probab, vol. 4, no. 3, pp. 485-510, Jul. 1991.

No Free-Lunch Theorems

- D. H. Wolpert, "The Lack of A Priori Distinctions Between Learning Algorithms," Neural Computation, vol. 8, no. 7, pp. 1341-1390, Oct. 1996.
- D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 67-82, Apr. 1997.

Bibliography

Generalization for Deep Neural Networks (recent works)

- B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks," arXiv:1707.09564 [cs], Jul. 2017.
- W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz, "Compressibility and Generalization in Large-Scale Deep Learning," arXiv:1804.05862 [cs, stat], Apr. 2018.
- B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring Generalization in Deep Learning," arXiv:1706.08947 [cs], Jun. 2017.
- P. Bartlett, D. J. Foster, and M. Telgarsky, "Spectrally-normalized margin bounds for neural networks," arXiv:1706.08498 [cs, stat], Jun. 2017.
- M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," arXiv:1509.01240 [cs, math, stat], Sep. 2015.
- K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, "Generalization in Deep Learning," arXiv:1710.05468 [cs, stat], Oct. 2017.