Machine Learning

PAC-Learning

François Fleuret

May 11th, 2011

Introduction
Classification

The usual setting for learning for classification:

- A training set,
- a family of classifiers,
- a test set.

Learning means to choose a classifier according to its performances on the training set to get good performances on the test set.

Introduction
Topic of this lecture

The goal of this lecture is to give an intuitive understanding of the Probably Approximately Correct learning (PAC learning for short) theory.

- Concentration inequalities,
- basic PAC results,
- relation with Occam's principle,

Figures are supposed to help. If they do not, ignore them.

We will use the following notation:

- $\mathcal{X}$ the space of the objects to classify (for instance images),
- $\mathcal{C}$ the family of classifiers,
- $S = ((X_1,\ Y_1),\ \ldots,\ (X_{2N},\ Y_{2N}))$ a random variable on $(\mathcal{X} \times \{0, 1\})^{2N}$ standing for the training and test samples,
- F a random variable on $\mathcal{C}$ standing for the learned classifier. It can be a deterministic function of S or not.

Introduction
Remarks

- The set $\mathcal{C}$ contains all the classifiers obtainable with the learning algorithm.

  For an ANN for instance, there is one element of $\mathcal{C}$ for every single configuration of the synaptic weights.

- The variable S is not one sample, but a family of $2N$ samples with their labels. It contains both the training and the test set.

Gap between training and test error

One fixed f

For every $f \in \mathcal{C}$, let $\xi(f, S)$ denote the difference between the training and the test errors of f estimated on $S = ((X_1, Y_1), \dots, (X_{2N}, Y_{2N}))$.

$$\xi(f, S) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{f(X_{N+i}) \neq Y_{N+i}\}}_{\text{test error}} - \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{f(X_i) \neq Y_i\}}_{\text{training error}}$$

Where $\mathbf{1}\{t\}$ is equal to 1 if t is true, and 0 otherwise. Since S is random, this is a random quantity.

Gap between the test and the training error

Data-dependent f

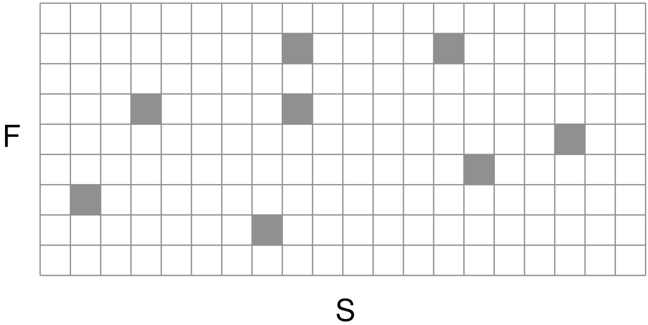
Given η, we want to bound the probability that the test error is less than the training error plus η.

$$P\left(\xi(F, S) \leq \eta\right) \quad \geq \quad ?$$

Here F is not constant anymore and depends on the $X_1, \ldots, X_{2N}$ and the $Y_1, \ldots, Y_N$.

Do figures help ?

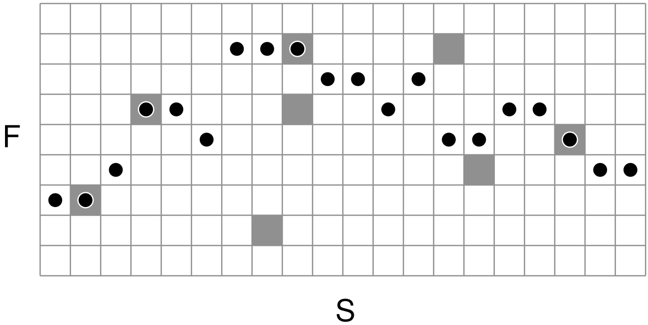
Violations of the error gap



Each row corresponds to a classifier, each column to a pair training/test set. Gray squares indicate $\xi(F, S) > \eta$.

Do figures help ?

A training algorithm



A training algorithm associates an *F* to every *S*, here shown with dots. We want to bound the number of dots on gray cells.

Where we see that for any fixed f, the test and training errors are likely to be similar . . .

Concentration Inequality

Hœffding's inequality (1963)

Given a family of independent random variables $Z_1, \ldots, Z_N$, bounded $\forall i, Z_i \in [a_i, b_i]$, if S denotes $\sum_i Z_i$, we have Hœffding's inequality (1963).

$$P(S - E(S) > t) \quad \leq \quad \exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right)$$

Concentration Inequality
Hœffding's inequality (1963)

Given a family of independent random variables $Z_1, \ldots, Z_N$, bounded $\forall i, Z_i \in [a_i, b_i]$, if S denotes $\sum_i Z_i$, we have Hœffding's inequality (1963).

$$P(S - E(S) > t) \quad \leq \quad \exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right)$$

This is an concentration result: It tells how much S is concentrated around its average value.

Concentration Inequality
Application to the error

Note that the $\mathbf{1}\{f(X_i) \neq Y_i\}$ are i.i.d Bernoulli, and we have

$$\begin{aligned}
\xi(f, S) &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{f(X_{N+i}) \neq Y_{N+i}\} - \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{f(X_i) \neq Y_i\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbf{1}\{f(X_{N+i}) \neq Y_{N+i}\} - \mathbf{1}\{f(X_i) \neq Y_i\}}_{\Delta_i}
\end{aligned}$$

Concentration Inequality
Application to the error

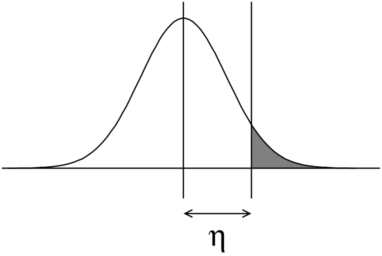
Note that the $\mathbf{1}\{f(X_i) \neq Y_i\}$ are i.i.d Bernoulli, and we have

$$\begin{aligned} \xi(f, S) &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{f(X_{N+i}) \neq Y_{N+i}\} - \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{f(X_i) \neq Y_i\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbf{1}\{f(X_{N+i}) \neq Y_{N+i}\} - \mathbf{1}\{f(X_i) \neq Y_i\}}_{\Delta_i} \end{aligned}$$

Thus ξ is the averaged sum of the Δ_i, which are i.i.d random variables on $\{-1, 0, 1\}$ of zero mean.

Concentration Inequality
Application to the error

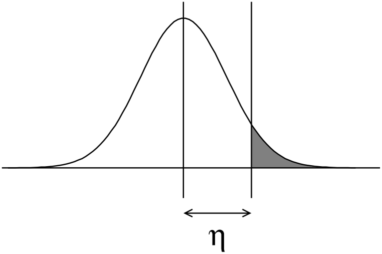
Hence, when f is fixed we have (Hœffding):



$$\forall f,\ \forall \eta,\quad P\left(\xi(f, S) > \eta\right) \quad \leq \quad \exp\left(-\frac{1}{2}\,\eta^2\,N\right)$$

Concentration Inequality
Application to the error

Hence, when f is fixed we have (Hœffding):



$$\forall f,\ \forall \eta,\quad P\left(\xi(f, S) > \eta\right) \quad \leq \quad \exp\left(-\frac{1}{2}\,\eta^2\,N\right)$$

(On our graph, we have an upper bound on the number of gray cells per row.)

Where we realize that the probability the chosen F *fails is lower than the probability that there exists a* f *that fails . . .*

Union bound

A first generalization bound

We have

$$P(\xi(F, S) > \eta) \quad = \quad \sum_{f} P(F = f, \xi(F, S) > \eta)$$

Union bound

A first generalization bound

We have

$$\begin{aligned} P(\xi(F, S) > \eta) &= \sum_f P(F = f, \xi(F, S) > \eta) \\ &= \sum_f P(F = f, \xi(f, S) > \eta) \end{aligned}$$

Union bound
A first generalization bound

We have

$$\begin{aligned} P(\xi(F, S) > \eta) &= \sum_f P(F = f, \xi(F, S) > \eta) \\ &= \sum_f P(F = f, \xi(f, S) > \eta) \\ &\leq \sum_f P(\xi(f, S) > \eta) \end{aligned}$$

Union bound

A first generalization bound

We have

$$\begin{aligned}
P(\xi(F, S) > \eta) &= \sum_f P(F = f, \xi(F, S) > \eta) \\
&= \sum_f P(F = f, \xi(f, S) > \eta) \\
&\leq \sum_f P(\xi(f, S) > \eta) \\
&\leq \|\mathcal{C}\| \exp\left(-\frac{1}{2}\eta^2 N\right)
\end{aligned}$$

Union bound

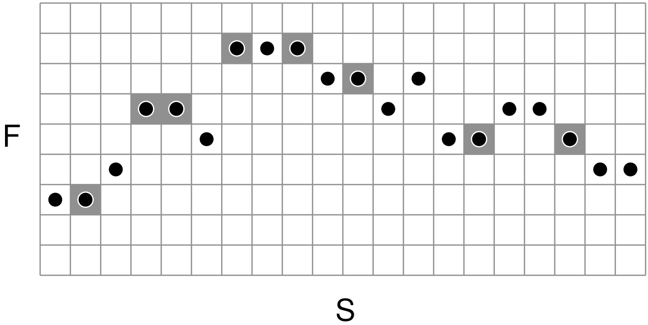
A first generalization bound

We have

$$\begin{aligned} P(\xi(F, S) > \eta) &= \sum_f P(F = f, \xi(F, S) > \eta) \\ &= \sum_f P(F = f, \xi(f, S) > \eta) \\ &\leq \sum_f P(\xi(f, S) > \eta) \\ &\leq \|\mathcal{C}\| \exp\left(-\frac{1}{2}\eta^2 N\right) \end{aligned}$$

This is our first generalization bound!

Do figures help ?
The union bound



We can see that graphically as a situation when the dots meet all the gray squares.

Union bound

We can fix the probability

If we define

$$\epsilon^\star \;=\; \|\mathcal{C}\| \,\exp\left(-\frac{1}{2}\,\eta^2\,N\right)$$

We have

$$\sqrt{2\,\frac{\log\|\mathcal{C}\| - \log\epsilon^\star}{N}} \;=\; \eta$$

Union bound

We can fix the probability

Hence from

$$P(\xi(F, S) > \eta) \quad \leq \quad \|\mathcal{C}\| \exp\left(-\frac{1}{2}\eta^2 N\right)$$

we get

$$P\left(\xi(F, S) > \sqrt{2\,\frac{\log\|\mathcal{C}\| + \log\frac{1}{\epsilon^\star}}{N}}\right) \quad \leq \quad \epsilon^\star$$

Union bound

We can fix the probability

Hence from

$$P(\xi(F, S) > \eta) \quad \leq \quad \|\mathcal{C}\| \exp\left(-\frac{1}{2}\eta^2 N\right)$$

we get

$$P\left(\xi(F, S) > \sqrt{2\,\frac{\log\|\mathcal{C}\| + \log\frac{1}{\epsilon^\star}}{N}}\right) \quad \leq \quad \epsilon^\star$$

Thus, with probability $1 - \epsilon^\star$, we know that the gap between the train and test error grows like the square root of the log of the number of classifiers $\|\mathcal{C}\|$.

Where we realize that we can arbitrarily distribute allowed errors on the fs before looking at the training data . . .

Prior on $\mathcal{C}$
What do we control

At that point, the only quantity we control is $\|\mathcal{C}\|$.

If we *know* that some of the mappings can be removed without hurting the train error, we can remove them and get a better bound.

Can we do something better than that?

Prior on $\mathcal{C}$

What do we control

At that point, the only quantity we control is $\|\mathcal{C}\|$.

If we *know* that some of the mappings can be removed without hurting the train error, we can remove them and get a better bound.

Can we do something better than that?

We introduce $\eta(f)$ as the control we want between the train and test error if f is chosen. Until now, this was constant.

Prior on $\mathcal{C}$

Let make η depend on F

Let $\epsilon(f)$ denote the (bound on the) probability that the constraint is not verified for f

$$\begin{aligned} P(\xi(F, S) > \eta(F)) &\leq P\left(\exists f \in \mathcal{C}, \xi(f, S) > \eta(f)\right) \\ &\leq \sum_f P(\xi(f, S) > \eta(f)) \\ &\leq \sum_f \epsilon(f) \end{aligned}$$

and we have

$$\forall f,\ \eta(f) = \sqrt{2 \frac{\log \frac{1}{\epsilon(f)}}{N}}$$

Prior on $\mathcal{C}$

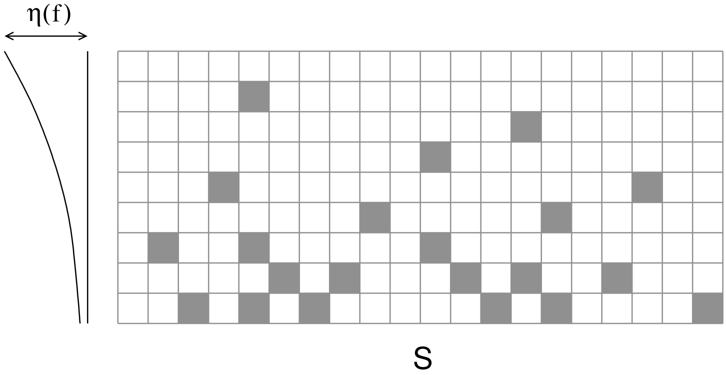
Let make η depend on F

Let define $\epsilon^\star = \sum_f \epsilon(f)$ and $\rho(f) = \frac{\epsilon(f)}{\epsilon^\star}$. The later is a distribution on $\mathcal{C}$.

Note that both can be fixed arbitrarily, and we have

$$\forall f,\ \eta(f) \quad = \quad \sqrt{2\,\frac{\log\frac{1}{\rho(f)} + \log\frac{1}{\epsilon^\star}}{N}}$$

Do figures help ?
When η depends on f



If the margin η depends on F, the proportion of gray squares is not the same on every row.

Prior on $\mathcal{C}$

Let's put everything together

Our final result is that, if

- we choose a distribution ρ on $\mathcal{C}$ arbitrarily,
- we choose $0 < \epsilon^\star < 1$ arbitrarily,
- we sample a pair S training set / test set each of size N,
- we choose a F after looking at the training set.

Prior on $\mathcal{C}$

Let's put everything together

Our final result is that, if

- we choose a distribution ρ on $\mathcal{C}$ arbitrarily,
- we choose $0 < \epsilon^\star < 1$ arbitrarily,
- we sample a pair S training set / test set each of size N,
- we choose a F after looking at the training set.

Then, we have with probability greater than $1 - \epsilon^\star$:

$$\xi(F, S) \quad \leq \quad \sqrt{2 \, \frac{\log \frac{1}{\rho(F)} + \log \frac{1}{\epsilon^\star}}{N}}$$

where $\xi(F, S)$ is the difference between the test and train errors.

Prior on $\mathcal{C}$

This is a philosophical theorem!

If we see $-\log \rho(f)$ as the "description" length of f (think Huffman).
Our result true with probability $\epsilon^\star$

$$\xi(F, S) \quad \leq \quad \sqrt{2\,\frac{\log \frac{1}{\rho(F)} + \log \frac{1}{\epsilon^\star}}{N}}$$

says that picking a classifier with a long description leads to a bad control on the test error.

Prior on $\mathcal{C}$

This is a philosophical theorem!

If we see $-\log \rho(f)$ as the "description" length of f (think Huffman). Our result true with probability $\epsilon^\star$

$$\xi(F, S) \quad \leq \quad \sqrt{2\,\frac{\log \frac{1}{\rho(F)} + \log \frac{1}{\epsilon^\star}}{N}}$$

says that picking a classifier with a long description leads to a bad control on the test error.

Entities should not be multiplied unnecessarily.

Principle of parsimony of William of Occam (1280 – 1349). Also known as Occam's Razor.

The end