Machine Learning
Statistical Learning Theory

Gerard Pons-Moll

06.02.2019
Program of today

A brief overview of results from statistical learning theory

- stochastic convergence,
- different notions of consistency,
- consistency for finite function classes,
- consistency for infinite function classes and the VC dimension,
- universal Bayes consistency - conditions?
- negative results: no free lunch theorem.
Motivation
Can we upper bound the deviation of $R(f_n)$ from
- the Bayes risk $R^* = \inf_{f \text{ measurable}} R(f)$
- the best risk $R_F = \inf_{f \in F} R(f)$ in the class $F$.

where $f_n$ is the function chosen by the learning algorithm.
Here: Binary classification, canonical zero-one loss.

Concentration
A random variable $X$ is concentrated if its distribution is very peaked around the expectation $\mathbb{E}X$ of $X$.

**empirical mean:** $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, with the $\{X_i\}_{i=1}^{n}$ i.i.d. sample.
Intuition: the distribution of $X$ will be concentrated around the true mean $\mathbb{E}X = \mathbb{E}X$. 
Three different notions of convergence of random variables

Definition

Let \( \{X_n\}, \ n = 1, 2, \ldots, \) be a sequence of random variables. We say that \( X_n \) **converges in probability**, \( \lim_{n \to \infty} X_n = X \) in probability, if for each \( \varepsilon > 0, \)

\[
\lim_{n \to \infty} P(|X_n - X| \geq \varepsilon) = 0.
\]

We say that \( X_n \) **converges almost surely (with probability 1)**, \( \lim_{n \to \infty} X_n = X \) almost surely (a.s.), if

\[
P(\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)) = 1.
\]

For a fixed \( p \geq 1 \) we say that \( X_n \) **converges in \( L_p \) or the \( p \)-th mean**, \( \lim_{n \to \infty} X_n = X \) in \( L_p \), if

\[
\lim_{n \to \infty} \mathbb{E}(|X_n - X|^p) = 0.
\]
Proposition

The following implications hold,

- \( \lim_{n \to \infty} \mathbb{E}(|X_n - X|^p) = 0 \implies P(|X_n - X| \geq \varepsilon) = 0, \)
- \( \lim_{n \to \infty} X_n = X \text{ almost surely } \implies P(|X_n - X| \geq \varepsilon) = 0, \)
- If for each \( \varepsilon > 0, \)

\[
\sum_{n=0}^{\infty} P(|X_n - X| \geq \varepsilon) < \infty,
\]

then \( \lim_{n \to \infty} X_n = X \text{ almost surely.} \)

Relevance for machine learning?

\( R(f_n) \) is a random variable since it depends on the training sample.

- how far is \( R(f_n) \) away from the Bayes risk \( R^* \)?
- In which sense \( \lim_{n \to \infty} R(f_n) = R^* \)?
Consistency (Classification)

Consistency for binary classification:
- Loss function, is 0-1-loss,
- \( R(f) = \mathbb{E} \mathbb{1}_{f(X) \neq Y} = P(f(X) \neq Y) \),
- Bayes risk \( R^* = \inf_{f \text{ measurable}} R(f) \).
- best risk in function class \( R_F = \inf_{f \in \mathcal{F}} R(f) \) in the class \( \mathcal{F} \).

Definition (Consistency)

A classification rule is
- consistent for a distribution of \((X, Y)\) if \( \lim_{n \to \infty} R(f_n) = R_F \),
- Bayes consistent for a distribution of \((X, Y)\) if \( \lim_{n \to \infty} R(f_n) = R^* \).

We have **weak** (convergence in probability) and **strong** (almost sure convergence) consistency.

The probability \( P(\left| R(f_n) - R^* \right| > \varepsilon) \) is with respect to all possible training samples of size \( n \).
What does consistency mean?

- The true error of $f_n$ converges to the best possible error,
- asymptotic property - no finite sample statements,
- distribution dependent, for example hard margin SVM’s are Bayes consistent for distributions where the support of $P(X|Y = 1)$ and $P(X|Y = -1)$ is linearly separable, but clearly for no problem which is non-separable.

A priori we should make no/too many assumptions about the true nature of the problem!
Definition (Universal consistency)

A classification rule/learning algorithm is **universally (weakly/strongly) consistent** if it is (weakly/strongly) consistent for any distribution on $\mathcal{X} \times \mathcal{Y}$.

- strong requirement, since the distribution might be arbitrarily strange.
- nevertheless there exist several universally consistent learning algorithms.

**Our main interest: universal consistency**
Consistency

Find the best possible function in a class of functions
Every learning algorithm selects either implicitly or explicitly the classifier $f_n$ from some function class $\mathcal{F}$,

Natural decomposition (bias-variance decomposition),

$$R(f_n) - R^* = R(f_n) - \inf_{f \in \mathcal{F}} R(f) + \inf_{f \in \mathcal{F}} R(f) - R^* .$$

- The estimation error is random since it depends on $f_n$ and thus on the training data - measures the deviation from the best possible risk in the hypothesis class $\mathcal{F}$.
- The approximation error is deterministic and measures the deviation of $R_\mathcal{F}$ from the Bayes risk $R^*$. It depends on the hypothesis class $\mathcal{F}$ and the data-generating measure - can only be bounded by making assumptions on the distribution of the data.
Downside of simple function classes

In the worst case we have $R^* = 0$ but $\inf_{f \in \mathcal{F}} R(f) \gg 0$.

**The XOR Problem**

\[
\begin{array}{cc}
Y=0 & Y=1 \\
Y=1 & Y=0
\end{array}
\]

*Figure: XOR-problem in $\mathbb{R}^2$. Linear classifiers $\mathcal{F} = \{ f(x) = \langle w, x \rangle + b \mid w \in \mathbb{R}^2, b \in \mathbb{R} \}$ are very bad but $R^* = 0$.***
Let $f_n$ be chosen by empirical risk minimization, that is $f_n = \arg \min_{f \in \mathcal{F}} R_n(f)$, where $R_n(f) = \frac{1}{n} \sum_{i=1}^{n} 1_{f(X_i) \neq Y_i}$. Then

$$R(f_n) - \inf_{f \in \mathcal{F}} R(f) \leq 2 \sup_{f \in \mathcal{F}} |R(f) - R_n(f)|.$$  

**Proof:** We have with $f^*_F = \arg \min_{f \in \mathcal{F}} R(f)$,

$$R(f_n) - \inf_{f \in \mathcal{F}} R(f) = R(f_n) - R_n(f_n) + R_n(f_n) - R(f^*_F) \leq R(f_n) - R_n(f_n) + R_n(f^*_F) - R(f^*_F) \leq 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|,$$

where the second inequality follows from the fact that $f_n$ minimizes the empirical risk.
Empirical Processes

Definition of empirical processes

Definition

A **stochastic process** is a collection of random variables \( \{Z_n, n \in T\} \) on the same probability space, indexed by an arbitrary index set \( T \). An **empirical process** is a stochastic process based on a random sample.

In statistical learning theory we are studying the empirical process,

\[
\sup_{f \in \mathcal{F}} |R_n(f) - R(f)|,
\]

since uniform control of the deviation \( R_n(f) - R(f) \) yields consistency!

\[
R(f_n) - \inf_{f \in \mathcal{F}} R(f) \leq 2 \sup_{f \in \mathcal{F}} |R(f) - R_n(f)|.
\]
Hoeffding’s inequality

**Theorem**

Let $X_1, \ldots, X_n$ be independent, bounded and identically distributed random variables such that $X_i$ falls in the interval $[a_i, b_i]$ with probability one. Then for any $\varepsilon > 0$ we have

$$
P\left(\left|\frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}X_i\right| \geq \varepsilon \right) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{\frac{1}{n} \sum_{i=1}^{n} (b_i - a_i)^2}\right).
$$

Control of the deviation for a **fixed** function with $R(f) = \mathbb{E}[\mathbb{1}_{f(X) \neq Y}]$,

$$
P\left(\left|R_n(f) - R(f)\right| \geq \varepsilon \right) \leq 2 \exp\left(-2n\varepsilon^2\right).
$$

**Important:** This cannot be simply applied to $f_n$ - the function found by empirical risk minimization - since $f_n$ depends on the training data.
Bounds for the case of a finite set of functions $\mathcal{F}$

**Proposition**

Let $\mathcal{F}$ be a finite set of functions, then

\[
P\left( \sup_{f \in \mathcal{F}} \left| R_n(f) - R(f) \right| \geq \varepsilon \right) \leq 2|\mathcal{F}| \exp\left( -2n\varepsilon^2 \right),
\]

where $|\mathcal{F}|$ is the cardinality of $\mathcal{F}$. And thus with probability $1 - \delta$,

\[
R(f_n) \leq R(f^*_\mathcal{F}) + \sqrt{\frac{\log |\mathcal{F}| + \log \frac{2}{\delta}}{n}}.
\]

**Proof:** Noting that $0 \leq 1_{f(X) \neq Y} \leq 1$ we get the result using Hoeffding’s inequality. Then with $\delta = 2|\mathcal{F}|e^{-2n\varepsilon^2}$ one gets $\varepsilon = \sqrt{\frac{1}{n}\left( \log |\mathcal{F}| + \log \frac{2}{\delta} \right)}$.

The **convergence rate** is of order $\frac{1}{\sqrt{n}}$ $\Rightarrow$ typical in SLT.
Infinite number of functions

Major contribution of Vapnik and Chervonenkis: uniform deviation bounds over general infinite classes.

Given points $x_1, \ldots, x_n$ and a class $\mathcal{F}$ of binary-valued functions denote by

$$\mathcal{F}_{x_1, \ldots, x_n} = \left\{ \{f(x_1), \ldots, f(x_n)\} \mid f \in \mathcal{F} \right\},$$

the set of all possible classification of the set of points via functions in $\mathcal{F}$.

**Definition**

The **growth function** $S_{\mathcal{F}}(n)$ is the maximum number of ways into which $n$ points can be classified by the function class $\mathcal{F}$,

$$S_{\mathcal{F}}(n) = \sup_{(x_1, \ldots, x_n)} |\mathcal{F}_{x_1, \ldots, x_n}|.$$

If $S_{\mathcal{F}}(n) = 2^n$ we say that $\mathcal{F}$ **shatters** $n$ points.
Why is this growth function interesting?

**Symmetrization lemma**

- **ghost sample**: a second i.i.d. sample of size $n$ (independent of the training data).
- $R'_n(f)$ denotes the empirical risk associated with the ghost sample.

**Lemma**

Let $n \varepsilon^2 \geq 2$, we have

$$P\left( \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| > \varepsilon \right) \leq 2 P\left( \sup_{f \in \mathcal{F}} |R_n(f) - R'_n(f)| > \frac{\varepsilon}{2} \right),$$

- **Important**: $|R_n(f) - R'_n(f)|$ depends only on the values of the function takes on the $2n$ samples - these are maximum $2^{2n}$ different values $\implies$ independent of how many functions are contained in $\mathcal{F}$.
- a simple union bound will now yield the $V(apnik)C(hervonenkis)$-bound.
The growth function is a measure of the “size” of $\mathcal{F}$.

**Theorem (Vapnik-Chervonenkis)**

For any $\delta > 0$, with probability at least $1 - \delta$,

$$
R(f_n) \leq R(f^*_\mathcal{F}) + 8 \sqrt{\frac{\log S_{\mathcal{F}}(2n) + \log \frac{8}{\delta}}{2n}}.
$$

**Proof:**

$$
P\left( R(f_n) - \inf_{f \in \mathcal{F}} R(f) > \varepsilon \right) \leq P\left( \sup_{f \in \mathcal{F}} |R(f) - R_n(f)| > \frac{\varepsilon}{2} \right)
\leq 2 P\left( \sup_{f \in \mathcal{F}} |R_n(f) - R'_n(f)| > \frac{\varepsilon}{4} \right)
\leq 2 S_{\mathcal{F}}(2n) P\left( |R_n(f) - R'_n(f)| > \frac{\varepsilon}{4} \right)
\leq 4 S_{\mathcal{F}}(2n) P\left( |R_n(f) - R(f)| > \frac{\varepsilon}{8} \right) \leq 8 S_{\mathcal{F}}(2n) e^{-\frac{n\varepsilon^2}{32}}.
$$
Discussion of VC-Bound

For a finite class \( \log S_{\mathcal{F}}(n) \leq |\mathcal{F}| \Rightarrow \) up to constants at least as good as the previous bound for finite \( \mathcal{F} \).

**Definition**

The **VC dimension** \( VC(\mathcal{F}) \) of a class \( \mathcal{F} \) is the largest \( n \) such that \( S_{\mathcal{F}}(n) = 2^n \).

What happens if \( \mathcal{F} \) can always realize all \( 2^n \) possibilities?

\[
R(f_n) \leq R(f^*_\mathcal{F}) + 8 \sqrt{\frac{\log S_{\mathcal{F}}(2n) + \log \frac{8}{\delta}}{2n}}
\]

\[
\leq R(f^*_{\mathcal{F}}) + 8 \sqrt{\frac{n \log 2 + \log \frac{8}{\delta}}{2n}}
\]

The second term does not converge to zero as \( n \to \infty \)!

\[\implies \] bound suggests that restricted \( \mathcal{F} \) is required for generalization.
What happens with $S_{\mathcal{F}}(n)$ for $n > \text{VC}(\mathcal{F})$?

We know: $n \leq \text{VC}(\mathcal{F}) \implies S_{\mathcal{F}}(n) = 2^n$ but what if $n > \text{VC}(\mathcal{F})$?

**Lemma (Vapnik-Chervonenkis, Sauer, Shelah)**

Let $\mathcal{F}$ be a class of functions with finite VC-dimension $\text{VC}(\mathcal{F})$. Then for all $n \in \mathbb{N}$,

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^{\text{VC}(\mathcal{F})} \binom{n}{i},$$

and for all $n > \text{VC}(\mathcal{F})$,

$$S_{\mathcal{F}}(n) \leq \left( \frac{e \cdot n}{\text{VC}(\mathcal{F})} \right)^{\text{VC}(\mathcal{F})}.$$ 

Phase transition from exponential to polynomial growth of $S_{\mathcal{F}}(n)$
Plugging the bounds on the growth function into the VC bounds

**Corollary**

Let $\mathcal{F}$ be a function class with VC-dimension $\text{VC}(\mathcal{F})$, then for $2n > \text{VC}(\mathcal{F})$ one has for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(f_n) \leq R(f^*_\mathcal{F}) + 8\sqrt{\text{VC}(\mathcal{F}) \log \frac{2e n}{\text{VC}(\mathcal{F})} + \log \frac{8}{\delta}} \frac{1}{2n}.$$ 

Deviation of $R(f_n)$ from $R(f^*_\mathcal{F}) = \inf_{f \in \mathcal{F}} R(f)$ decays as $\sqrt{\text{VC}(\mathcal{F}) \log \frac{n}{n}}$.

- VC dimension is not just counting the number of functions but the variability of the functions in the class on the sample.
- finite VC dimension ensures **universal consistency**,
- other techniques for bounds exist: covering numbers, Rademacher averages.
Necessary and sufficient conditions for consistency
The following theorem is one of the key-theorems for statistical learning.

Theorem (Vapnik-Chervonenkis (1971))

A necessary and sufficient condition for the universal consistency of empirical risk minimization using a function class $\mathcal{F}$ is,

$$
\lim_{n \to \infty} \frac{\log S_{\mathcal{F}}(n)}{n} = 0
$$

We have proven that $\lim_{n \to \infty} \frac{\log S_{\mathcal{F}}(n)}{n} = 0$ is sufficient for consistency. The proof, that this condition is also necessary requires a bit more effort.
Is the restriction necessary?

Empirical risk minimization can be inconsistent

Input space: \( \mathcal{X} = [0, 1] \). The labels are deterministic

\[
Y = \begin{cases}
  -1, & \text{if } X \leq 0.5, \\
  1, & \text{if } X > 0.5.
\end{cases}
\]

and \( P(X \leq 0.5) = \frac{1}{2} \).

We consider the following classifier,

\[
f_n(X) = \begin{cases}
  Y_i & \text{if } X = X_i \text{ for some } i = 1, \ldots, n \\
  1 & \text{otherwise}.
\end{cases}
\]

We have \( R_n(f_n) = 0 \) but \( R(f_n) = \frac{1}{2} \).

The classifier \( f_n \) is **not Bayes consistent**. We have,

\[
\lim_{n \to \infty} R(f_n) = \frac{1}{2} \neq 0 = R^*.
\]

\[\implies\] just memorizing - no learning, no generalization.
VC dimensions of selected function classes:

- The set of linear halfspaces in $\mathbb{R}^d$ has VC dimension $d + 1$.
- The set of linear halfspaces of margin $\rho$ and where the smallest sphere enclosing the data has radius $R$ has VC dimension,

$$VC(\mathcal{F}) \leq \min \left\{ d, \frac{4 R^2}{\rho^2} \right\} + 1.$$

- The function $\text{sign}(\sin(tx))$ on $\mathbb{R}$ has infinite VC dimension.

$\Rightarrow$ VC dimension has nothing to do with the number of free parameters!
VC Bounds and SVM

Justification for Support Vector machines

The set of linear halfspaces of margin $\rho$ and where the smallest sphere enclosing the data has radius $R$ has VC dimension,

$$VC(\mathcal{F}) \leq \min \left\{ d, \frac{4 R^2}{\rho^2} \right\} + 1.$$  

The vector $w$ of the optimal maximal-margin hyperplane satisfies,

$$\|w\|^2 = \frac{1}{\rho^2},$$

Thus, the Support-Vector Machine (SVM)

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - Y_i(\langle w, X_i \rangle + b)\} + \lambda \|w\|^2.$$  

penalizes large margins $\|w\| \implies \textbf{limits capacity of function class}$
Remarks on VC bounds (applies also to other existing bounds)

- **No a-posteriori justification**: bounds cannot be used for a posteriori justification. In particular, the bound holds not for the margin obtained by the SVM, but the bound holds for a function class with pre-defined margin (before seeing the data)!

- **Bounds are often loose**: the bounds are worst-case bounds which apply to any possible probability measure on $\mathcal{X} \times \mathcal{Y} \implies$ for practical sample sizes bounds are often larger than 1! But: bounds capture certain characteristics of the learning algorithm.
Universal Bayes consistency

Decomposition into estimation and approximation error),

\[ R(f_n) - R^* = R(f_n) - \inf_{f \in \mathcal{F}} R(f) + \inf_{f \in \mathcal{F}} R(f) - R^* . \]

\hspace{1cm} \text{Estimation error} \quad \text{Approximation error}

\implies \text{up to now fixed function class} \implies \text{fixed approximation error.}

Structural risk minimization:

- Let the function class \( \mathcal{F} \) be a function of the sample size \( n \): \( \mathcal{F}_n \).
- as \( n \to \infty \) let \( \mathcal{F}_n \) grow so that in the limit it can model any function but estimation error is still bounded:

\[ \text{withprob.} \geq 1 - \delta, \quad R(f_n) \leq R(f^*_F) + 8\sqrt{\frac{\text{VC}(\mathcal{F}_n) \log \frac{2e^n}{\text{VC}(\mathcal{F}_n)} + \log \frac{8}{\delta}}{2n}}. \]

\implies \text{Universal Bayes consistency}
Naturally arising questions

- Can we quantify the convergence to the Bayes risk? Can we obtain rates of convergence?
- What does universal consistency mean for the finite sample case?
- Is there a universally best learning algorithm?
First negative result

Intuition: For every fixed $n$ there exists a distribution where the classifier is arbitrarily bad!

Theorem

For any $\varepsilon > 0$ and any integer $n$ and classification rule $f_n$, there exists a distribution of $(X, Y)$ with Bayes risk $R^* = 0$ such that

$$\mathbb{E}[R(f_n)] \geq \frac{1}{2} - \varepsilon.$$ 

- construct a distribution on the set $\mathcal{X} = \{1, \ldots, K\}$,
- noise-free but no structure,
- for fixed $n$ choose $K$ sufficiently large such that the rule $f_n$ will fail completely on the rest of $\mathcal{X}$. 
First negative result

There exists no universally consistent learning algorithm such that \( R(f_n) \) converges uniformly over all distributions to \( R^* \).
Second negative result

Theorem

Let \( \{a_n\} \) be a sequence of positive numbers converging to zero with
\[
\frac{1}{16} \geq a_1 \geq a_2 \geq \ldots.
\]
For every sequence of classification rules, there exists a distribution of \((X, Y)\) with \(R^* = 0\), such that for all \(n\),
\[
\mathbb{E}[R(f_n)] \geq a_n.
\]

This result states that universally good learning algorithms do not exist \(\Rightarrow\) convergence to the Bayes risk can be **arbitrarily slow**!

There exist no universal rates to the Bayes risk. If one wants to have rates of convergence to the Bayes risk one has to restrict the class of distributions on \(\mathcal{X} \times \mathcal{Y}\).
Third negative result

Theorem

For every sequence of classification rules $f_n$, there is a universally consistent sequence of classification rules $g_n$ such that for some distribution on $\mathcal{X} \times \mathcal{Y}$

$$P(f_n(X) \neq Y) > P(g_n(X) \neq Y), \quad \forall n \geq 0.$$ 

Thus for every universally consistent learning rule there exists a distribution on $\mathcal{X} \times \mathcal{Y}$ such that another universally consistent learning rule is strictly better.

There exists no universally superior learning algorithm.
Summary

1 Restriction of the class of distributions on $\mathcal{X} \times \mathcal{Y}$ $\implies$ convergence rates to Bayes for universally consistent learning algorithms.

**Problem:** Assumptions cannot be tested. Performance guarantees are only valid under the made assumptions.

2 Restriction of the function class $\implies$ no universal consistency possible.

Comparison to the best possible function in the class is possible uniformly over all distributions.

But **no performance guarantees** with respect to the Bayes risk.
Convergence rates to Bayes only possible under assumptions on the distribution of $(X, Y)$

Reasonable assumptions fulfill two requirements:

- The assumptions should be as natural as possible, meaning that one expects that most the data generating distributions one encounters in nature fulfill these assumptions.
- The assumptions should be narrow enough, so that one can still prove convergence rates.
Assumptions

In terms of the regression function: \( \eta(x) = \mathbb{E}[Y|X=x] \).
- \( \eta(x) \) lies in some Sobolev space (has certain smoothness properties),
- Margin/low noise conditions introduced by Massart and Tsybakov,

Definition

A distribution \( P \) on \( \mathcal{X} \times \{-1, 1\} \) fulfills the low noise condition if there exist constants \( C > 0 \) and \( \alpha \geq 0 \) such that

\[
P\left( |\eta(X)| \leq t \right) \leq Ct^\alpha, \quad \forall t \geq 0.
\]

The coefficient \( \alpha \) is called the noise coefficient of \( P \).

1. \( \alpha = 0 \) is trivial and implies no restrictions on the distribution,
2. \( \alpha = \infty, \eta(x) \) strictly bounded away from zero.
Universal consistency for soft-margin SVM’s

**Definition**
A continuous kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called **universal** if the associated RKHS $\mathcal{H}_k$ is dense in the set of continuous functions $C(X)$ with the $\| \cdot \|_\infty$-norm, that is for all $f \in C(X)$ and $\varepsilon > 0$ there exists a $g \in \mathcal{H}_k$ such that

$$\| f - g \|_\infty \leq \varepsilon.$$  

$\Rightarrow$ Measurable functions can be approximated by continuous functions.

A soft-margin SVM in $\mathbb{R}^d$ with a **universal kernel** is universally consistent.

**Theorem**
Let $\mathcal{X} \subset \mathbb{R}^d$ be compact, then the soft-margin SVM with error parameter $C_n = n^{1-\beta}$ for some $0 < \beta < \frac{1}{d}$ and a Gaussian kernel is universally...
Bachelor/Master/PhD topics in machine learning!

Thanks for your attention!