Statistical Learning Theory: the informal roadmap

- Started by Vapnik and Chervonenkis in the Sixties (Vapnik 1995, 1998)
- Statistical point of view: data generated by an unknown stochastic source
- Problem (supervised learning): how to guarantee that the empirical error converges to true error ?
  - Law of large numbers ? No, increasing the size of the training set is not sufficient by itself to guarantee convergence, we need to require a statistical property called *consistency*
  - a necessary and sufficient condition to guarantee *consistency* is *uniform convergence*
- It turns out that VC-dimension can be used to derive bounds on uniform convergence
- Support Vector Machines use hypotheses with “large margin” (small VC-dimension)

Probability Tools: basic facts

Let A and B be some events (i.e. elements of a σ-algebra), and X some real-values random variable.

- Basic Facts
  - Union: $\mathbb{P}[A \text{ or } B] \leq \mathbb{P}[A] + \mathbb{P}[B]$
  - Inclusion: if $A \Rightarrow B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$
  - Inversion: if $\mathbb{P}[X > t] \leq F(t)$ then with probability at least $1 - \delta$, $X \leq F^{-1}(\delta)$
  - Expectation: if $X \geq 0$,

$$E[X] = \int_0^{\infty} \mathbb{P}[X \geq t] dt$$

- Hoeffding: Let $X_1, \ldots, X_n$ be n i.i.d. random variables with $f(X) \in [a, b]$.Then $\forall \epsilon > 0$, we have

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E[f(X)]\right| > \epsilon\right] \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

Statistical Learning Theory: data and risk

- Training Set: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, generated i.i.d from $P(x, y)$
  - if $y_i \in \mathbb{R}$, then we have a *regression* task
  - if $y_i \in \{-1, 1\}$, then we have a (binary) *classification* task
- let focus on binary *classification* tasks
- in SLT the true error is called *risk* (or *expected loss*) and is written as

$$R[h] = \int Loss(x, y, h(x))\, dP(x, y)$$

  where $h()$ is a hypothesis (output in $\{-1, 1\}$) and $Loss()$ is a function which measures how much any specific error costs
- for now let consider $Loss(x, y, h(x)) = \frac{1}{2}|h(x) - y| \in \{0, 1\}$

Statistical Learning Theory: Induction Principle

The problem in finding

$$h^{opt} = arg \min_{h \in \mathcal{H}} R[h]$$

(notice that h^{opt} may not be unique) is that

- we don't know $P(x, y)$
- we just have a finite training set

How to use the training set ? We need to define an *Induction Principle*
We can minimize the training error (empirical error)

$$R_{emp}[h] = \frac{1}{n} \sum_{i=1}^{n} |h(x_i) - y_i|$$

This corresponds to use as *Induction Principle* the so called

Empirical Risk Minimization (ERM)

Statistical Learning Theory: Problem

The main problem: can we use the Law of Large Numbers to guarantee

$$R_{emp}[h] \rightarrow R[h] \text{ as } n \rightarrow \infty \text{ ?}$$

Let use a statistical point of view:

- let define $\xi_i = \frac{1}{2}|h(x_i) - y_i|$
- since all the examples are drawn independently, then we are faced with *Bernoulli trials*
- thus the $\xi_1, \ldots, \xi_n$ are independently sampled from a random variable defined as $\xi = \frac{1}{2}|h(x) - y|$

There is a famous inequality which characterizes how the empirical mean $\frac{1}{n}\sum_{i=1}^{n} \xi_i$ converges to the expected value (or expectation) of ξ, denoted by $\mathrm{E}[\xi]$:

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mathrm{E}[\xi]\right| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2}$$

Statistical Learning Theory: Law of Large Numbers

Great! Recalling that

- $R_{emp}[h] = \frac{1}{n}\sum_{i=1}^{n} \xi_i$
- $R[h] = \mathrm{E}(\xi)$

we get

$$\mathbb{P}\left[|R_{emp}[h] - R[h]| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2}$$

So, not only the empirical risk converges to the risk, but it converges exponentially fast in the number of training example!
WARNING: the bound is probabilistic in nature, i.e. it does not rule out the presence of cases where the deviation is large. However, if we have many hypotheses, the probability that

$$h^{\downarrow} = arg \min_{h \in \mathcal{H}} R_{emp}[h]$$

($h^{\downarrow}$ need not be unique) will have a large deviation seems to be very small...

Statistical Learning Theory: Law of Large Numbers

...however, $h^{\downarrow}$ is a very atypical hypothesis since it tries to reduce the mean of the ξ_i as small as possible, moving away from the "natural" average loss of the variable ξ

we are no longer looking at independent Bernoulli trials!

So the learning process (using the ERM) is looking for the **worst case**
In conclusion, the Law of Large Numbers by itself is not enough!

What we actually need is *consistency*, i.e.:

$$R_{emp}[h^{\downarrow}] \rightarrow R[h^{opt}] \quad \textbf{and} \quad R[h^{\downarrow}|Tr] \rightarrow R[h^{opt}] \textbf{ as } n \rightarrow \infty$$

where $R[h|Tr] = \sum_{i=1}^{n} \frac{1}{2}|h(x_i) - y_i|P(x_i, y_i)$**, i.e. the unknown true risk evaluated on the training set**

Actually we need *nontrivial consistency*: consistency should hold for ALL hypotheses...

Statistical Learning Theory: Uniform Convergence

Key Theorem of Learning Theory (Vapnik and Chervonenkis, 1989):
For bounded loss functions, the ERM principle is consistent if and only if the empirical risk (i.e., empirical error) *converges uniformly* to the (true) risk in the following sense:

$$\lim_{n\to\infty} \mathbb{P}[\sup_{h\in\mathcal{H}}(R[h] - R_{emp}[h]) > \epsilon] = 0$$

for all $\epsilon > 0$

Statistical Learning Theory: Bounds

The **Key Theorem of Learning Theory** tell us that we have to focus our attention to the following probability:

$$\mathbb{P}[\sup_{h \in \mathcal{H}}(R[h] - R_{emp}[h]) > \epsilon]$$

What we can do is to try to derive upper bounds that can be used to establish under which conditions increasing the size of the training set implies a significant reduction of the probability itself.
Two important "tools" for deriving such bounds are:

- the *Union Bound*
- *Symmetrization*

Statistical Learning Theory: the Union Bound

If we just have two hypothesis, h_1 and h_2, in our Hypothesis Space, then uniform convergence of risk trivially follows from the law of large numbers. In fact, let us define

$$C_\epsilon^i \equiv \{ Tr \mid (R[h_i] - R_{emp}[h_i]) > \epsilon \}$$

then, by definition

$$\begin{aligned}
\mathbb{P}[\sup_{h \in \mathcal{H}} (R[h] - R_{emp}[h]) > \epsilon] &= \mathbb{P}(C_\epsilon^1 \cup C_\epsilon^2) \\
&= \mathbb{P}(C_\epsilon^1) + \mathbb{P}(C_\epsilon^2) - \mathbb{P}(C_\epsilon^1 \cap C_\epsilon^2) \\
&\leq \mathbb{P}(C_\epsilon^1) + \mathbb{P}(C_\epsilon^2)
\end{aligned}$$

Statistical Learning Theory: the Union Bound

Generalizing to a **finite** set of hypotheses $\mathcal{H} \equiv \{h_i, \ldots, h_k\}$, we get the *Union Bound*:

$$\mathbb{P}[\sup_{h \in \mathcal{H}}(R[h] - R_{emp}[h]) > \epsilon] = \mathbb{P}(C_\epsilon^1 \cup \ldots \cup C_\epsilon^k) \leq \sum_{i=1}^k \mathbb{P}(C_\epsilon^i)$$

Finally, we apply the Law of Large Numbers for each *individual* $\mathbb{P}(C_\epsilon^i)$ and since we have a finite number of these terms, uniform convergence is guaranteed:

$$\mathbb{P}[\exists h \in \{h_1, \ldots, h_k\} : R[h] - R_{emp}[h]) > \epsilon] \leq \sum_{i=1}^{k} \mathbb{P}(C_\epsilon^i) \leq k e^{-2n\epsilon^2}$$

Thus, if the Hypothesis Space is finite, we can use the Union Bound... but what happens if the Hypothesis Space is infinite ? ...we end up having an infinite number of nonzero quantities!

Statistical Learning Theory: Symmetrization

Vapnik and Chervonenkis solved this problem by reducing the infinite case to the finite case, via the introduction of the so called *ghost sample*.

In few words: the probability that the empirical risk differs from the true risk by more than ϵ, can be bounded by twice the probability that it differs from the empirical risk on a *second* sample (*test* set) of the same size n by more then $\epsilon/2$

Symmetrization (Vapnik and Chervonenkis):
For $n\epsilon^2 \geq 2$, we have

$$\mathbb{P}[\sup_{h\in\mathcal{H}}(R[h] - R_{emp}[h]) > \epsilon] \leq 2\mathbb{P}[\sup_{h\in\mathcal{H}}(R_{emp}[h] - R'_{emp}[h]) > \epsilon/2]$$

Here, the first P refers to the distribution of i.i.d. samples of size n, while the second one refers to i.i.d. samples of size $2n$. In the latter case, R_{emp} measures the loss on the first half of the sample, and R'_{emp} on the second half.

Implication of Symmetrization

Symmetrization is telling us that, for the purpose of bounding, the Hypothesis Space can be considered finite:

the number of distinct (boolean) functions (recall we are considering binary classification) over $2n$ elements is 2^{2n}.

Let $Tr_{2n} \equiv \{(x_1, y_1), \ldots, (x_{2n}, y_{2n})\}$ denote the given $2n$-sample, and denote by $\mathcal{N}(\mathcal{H}, Tr_{2n})$ the number of hypothesis that can be distinguished from their values on $\{x_1, \ldots, x_{2n}\}$

Then we can characterize the *capacity* of an (infinite) Hypothesis Space $\mathcal{H}$ by looking at the maximum (over all possible choices of a $2n$-sample) number of distinct functions that can be implemented by $\mathcal{H}$, denoted as $\mathcal{N}(\mathcal{H}, 2n)$.

But, wait a moment! This looks familiar...it is connected with *shattering*!!

In fact, the function $\mathcal{N}(\mathcal{H}, n)$ is referred to as the *shattering coefficient*.

Uniform Convergence Bound

Now we are ready to derive a bound for uniform convergence.
Using symmetrization, we have to bound

$$\mathbb{P}[\sup_{h\in\mathcal{H}}(R_{emp}[h] - R'_{emp}[h]) > \epsilon/2]$$

The basic idea is as follows:

1. pick a maximal set of hypotheses $\{h_1, \dots, h_{\mathcal{N}(\mathcal{H}, Tr_{2n})}\}$ that can be distinguished based on their values on Tr_{2n}
2. then use the Union Bound
3. finally bound each term by the first bound we introduced

However, before doing this an auxiliary step of randomization should be performed since each h_i depends on Tr_{2n}.
We skip the technical proof...

Uniform Convergence Bound

... and go directly to the final result:

$$\begin{aligned} \mathbb{P}[\sup_{h\in\mathcal{H}}(R_{emp}[h] - R'_{emp}[h]) > \epsilon/2] &\leq 4\mathrm{E}[\mathcal{N}(\mathcal{H}, Tr_{2n})]e^{-n\epsilon^2/8} \\ &= 4e^{-\ln \mathrm{E}[\mathcal{N}(\mathcal{H}, Tr_{2n})]n\epsilon^2/8} \end{aligned}$$

We conclude that if $\mathrm{E}[\mathcal{N}(\mathcal{H}, Tr_{2n})]$ does not grow exponentially in n, then we get a nontrivial and potentially useful bound.
Similar bounds can be derived within the field of empirical processes (*concentration*).
The term $\ln\ \mathrm{E}[\mathcal{N}(\mathcal{H}, Tr_{2n})]$ (called *annealed entropy*) is difficult to evaluate (it depends on a possibly unknown distribution...)
Because of that, it is substituted by other capacity concepts, e.g.:

$$\ln\ \mathrm{E}[\mathcal{N}(\mathcal{H}, Tr_n)] \leq \ln\ \mathcal{N}(\mathcal{H}, n) \equiv \mathcal{G}_{\mathcal{H}}(n) \quad \text{(Growth function)}$$

Growth function and VC-dimension

It is not difficult to recognize that the Growth function $\mathcal{G}_{\mathcal{H}}(n)$ and VC-dimension are intimately related: the VC-dimension is the maximal number of instances which can be shattered by a Hypothesis Space $\mathcal{H}$.
Thus, if we study the behavior of the Growth function as a function of the sample size n, we get:

- if $n \leq VC(\mathcal{H})$ then $\mathcal{G}_{\mathcal{H}}(n) = n \ln(2)$ (useless for the bound)
- if $n > VC(\mathcal{H})$ then it is possible to prove that

$$\mathcal{G}_{\mathcal{H}}(n) \leq VC(\mathcal{H}) \left(\ln(\frac{n}{VC(\mathcal{H})}) + 1 \right)$$

and the bound becomes useful: learning can succeed!

Confidence Intervals

The uniform convergence bound can be rewritten in a "PAC-learning style" by specifying the probability with which we want the bound to hold, an then derive a confidence interval. This can be done as follows:

- set $\delta = 4e^{-\ln \mathrm{E}[\mathcal{N}(\mathcal{H}, Tr_{2n})] n\epsilon^2/8}$
- solve the above equation with respect to ϵ

The result is that, with probability at least $1 - \delta$

$$R[h] \leq R_{emp}[h] + \sqrt{\frac{8}{n}\left(\ln \mathrm{E}[\mathcal{N}(\mathcal{H}, Tr_{2n})] + \ln \frac{4}{\delta}\right)}$$

which holds $\forall h \in \mathcal{H}$, and in particular for the hypothesis $h^{\downarrow}$ minimizing the empirical risk
Using VC-dimension to bound the annealed entropy, the general structure of bound of this type is

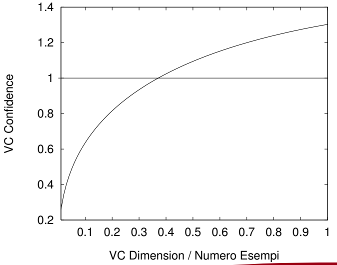
$$R[h] \leq \underbrace{R_{emp}[h]}_{A} + \underbrace{\epsilon(n, VC(\mathcal{H})/n, \delta)}_{B}$$

Confidence Intervals and VC-dimension

Where

- A ONLY DEPENDS on the hypothesis returned by the learning algorithm
- B is INDEPENDENT from the hypothesis returned by the learning algorithm, however it DEPENDS on the ratio between $VC(\mathcal{H})$ and the number of training examples n, and from the confidence $(1-\delta)$ with which the bound holds

B is usually called VC-confidence and it is monotone with respect to $\frac{VC(\mathcal{H})}{n}$; given n it grows with $VC(\mathcal{H})$.



Structural Risk Minimization

Problem: as the VC-dimension grows, the empirical risk (A) decreases, however the VC confidence (B) increases !

Because of that, Vapnik and Chervonenkis proposed a new inductive principle, i.e. Structural Risk Minimization (SRM), which aims to minimizing the right hand of the confidence bound, so to get a tradeoff between A and B:

Consider $\mathcal{H}_i$ such that

- $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_n$
- $VC(\mathcal{H}_1) \leq \cdots \leq VC(\mathcal{H}_n)$
- select the hypothesis with the smallest bound on the true risk

Example: Neural networks with an increasing number of hidden units