Statistical learning theory

«Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm, introduced in the 1920's and 1930's was replaced by a new one. This paradigm reflects a new answer to the fundamental question:

What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations?

CATS: ALL YOUR BASE ARE BELONG TO US.
ALL YOUR BAYES ARE BELONG TO US

In Fisher's paradigm the answer was very restrictive—one must know almost everything. [...]
The new paradigm overcame the restriction of the old one.»

Vladimir Vapnik
The Nature of Statistical Learning Theory (2000)

The formal setup of SLT

SLT deals mainly with **supervised learning** problems.

Given:

- ✓ an input (feature) space: $\mathcal{X}$
- ✓ an output (label) space: $\mathcal{Y}$ (typically $\mathcal{Y} = \{ -1, +1 \}$)

the question of learning amounts to estimating a functional relationship between the input and the output spaces:

$$f : \mathcal{X} \to \mathcal{Y}$$

Such a mapping f is called a **classifier**.

In order to do this, we have access to some (labeled) training data:

$$(X_1,Y_1), \dots , (X_n,Y_n) \in \mathcal{X} \times \mathcal{Y}$$

A **classification algorithm** is a procedure that takes the training data as input and outputs a classifier f.

Assumptions

In SLT one makes the following assumptions:

- ✓ there exists a joint probability distribution P on $\mathcal{X} \times \mathcal{Y}$
- ✓ the training examples (X_i, Y_i) are sampled independently from P (iid sampling).

In particular:

1. No assumptions on P
2. The distribution P is unknown at the time of learning
3. Non-deterministic labels due to label noise or overlapping classes
4. The distribution P is fixed

Losses and risks

We need to have some measure of "how good" a function f is when used as a classifier. A *loss function* measures the "cost" of classifying instance $X \in \mathcal{X}$ as $Y \in \mathcal{Y}$.

The simplest loss function in classification problems is the **0-1 loss** (or misclassication error):

$$\ell(X, Y, f(X)) = \begin{cases} 1 & \text{if } f(X) \neq Y \\ 0 & \text{otherwise.} \end{cases}$$

The *risk* of a function is the average loss over data points generated according to the underlying distribution P:

$$R(f) := E(\ell(X, Y, f(X)))$$

The *best classifier* is the one with the smallest risk $R(f)$.

Bayes classifiers

Among all possible classifiers, the “best” one is the *Bayes classifier*:

$$f_{Bayes}(x) := \begin{cases} 1 & \text{if } P(Y = 1 \mid X = x) \geq 0.5 \\ -1 & \text{otherwise.} \end{cases}$$

In practice, it is impossible to directly compute the Bayes classifier as the underlying probability distribution *P* is unknown to the learner.

The idea of estimating *P* from data doesn’t usually work …

Bayes’ theorem

> «[Bayes’ theorem] is to the theory of probability what Pythagoras’ theorem is to geometry.»
>
> Harold Jeffreys
> *Scientific Inference* (1931)

$$P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)} = \frac{P(e \mid h)P(h)}{P(e \mid h)P(h) + P(e \mid \neg h)P(\neg h)}$$

- ✓ *P*(*h*): prior probability of hypothesis *h*
- ✓ *P*(*h* | *e*): posterior probability of hypothesis *h* (in the light of evidence *e*)
- ✓ *P*(*e* | *h*): “likelihood” of evidence *e* on hypothesis *h*

The classification problem

Given:

- ✓ a set training points $(X_1,Y_1), \dots, (X_n,Y_n) \in \mathcal{X} \times \mathcal{Y}$ drawn iid from an *unknown* distribution P
- ✓ a loss functions

Determine a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ which has risk $R(f)$ as close as possible to the risk of the Bayes classifier.

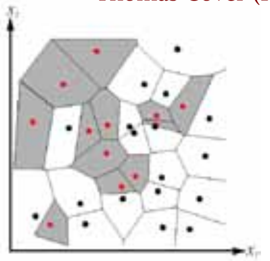
Caveat. Not only is it impossible to compute the Bayes error, but also the risk of a function f cannot be computed without knowing P.

A desperate situation?

An example: The nearest neighbor rule

> «Early in 1966 when I first began teaching at Stanford, a student, Peter Hart, walked into my office with an interesting problem. He said that Charles Cole and he were using a pattern classification scheme which, for lack of a better word, they described as the **nearest neighbor procedure**.
> This scheme assigned to an as yet unclassified observation the classification of the nearest neighbor. Were there any good theoretical properties of this procedure?»

Thomas Cover (1982)



… and its medieval origins

«Hence, when sight perceives some visible object, the faculty of discrimination immediately seeks its counterpart among the forms persisting in the imagination, and when it finds some form in the imagination that is like the form of that visible object, it will recognize that visible object and will perceive what kind of object it is.»

Alhazen, *The Books of Optics* (c. 1030)
(Eng. Transl.: M. Smith, 2001)

Arguably the first explicit formulation of the NN classification rule on record! (see: M. Pelillo, "Alhazen and the nearest neighbor rule," *Pattern Recognition Letters*, 2014).

How good is the NN rule?

Cover and Thomas showed that:

$$R(f_{Bayes}) \le R_\infty \le 2R(f_{Bayes})$$

where R_∞ denotes the expected error rate of NN when the sample size tends to infinity.

We cannot say anything stronger as there are probability distributions for which the performance of the NN rule achieves either the upper or lower bound.

Variations:

- ✓ ***k*-NN rule:** use the k nearest neighbors and take a majority vote
- ✓ **k_n-NN rule:** the same as above, for k_n growing with n

Theorem (Stone, 1977) If $n \to \infty$ and $k \to \infty$, such that $k/n \to 0$, then for all probability distributions $R(k_n\text{-NN}) \to R(f_{Bayes})$ (that is, the k_n-NN rule is "universally Bayes consistent").

Empirical Risk Minimization

«At the end of the 1960's, the theory of Empirical Risk Minimization (ERM) for the pattern recognition problem was constructed.

This theory included both (a) the general *qualitative theory* of generalization that described the necessary and sufficient conditions for consistency of the ERM induction principle [...];

and (b) the general *quantitative theory* that described the bounds on the probability of the (future) test error.»

Vladimir Vapnik
Statistical Learning Theory (1998)

The ERM principle

Instead of looking for a function which minimizes the true risk $R(f)$, we try to find one which minimizes the *empirical risk*:

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i, Y_i, f(X_i))$$

Given training data $(X_1,Y_1), \ldots , (X_n,Y_n) \in \mathcal{X} \times \mathcal{Y}$, a function space $\mathcal{F}$, and a loss function, we define the classifier f_n as:

$$f_n := \operatorname*{argmin}_{f\in\mathcal{F}} \mathrm{R}_{\text{emp}}(f)$$

This approach is called the *empirical risk minimization* (ERM) induction principle, the motivation of which comes from the law of large numbers.

Note. Same as least-squares/ML methods (but... binary *vs.* real functions!).

A key question

What has to be true of the function class $\mathcal{F}$ so that, no matter what the unknown background probability distribution, ERM eventually does as well as possible with respect to the rules in $\mathcal{F}$?

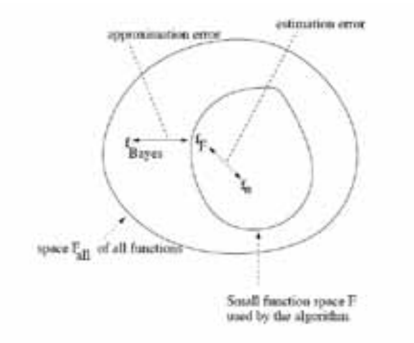
A fundamental result of SLT is that the set of rules in $\mathcal{F}$ cannot be too rich, where the richness of $\mathcal{F}$ is measured by its “VC dimension”.

Estimation *vs.* approximation

Ideally we want to make $R(f_n) - R(f_{Bayes})$ as small as possible, as $n \to \infty$.

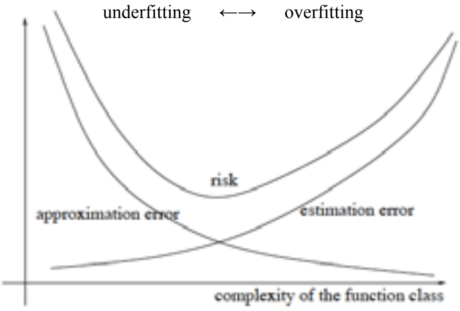
Denoting by $f_{\mathcal{F}}$ the best classifier in $\mathcal{F}$, the difference can be decomposed as:

$$R(f_n) - R(f_{Bayes}) = \underbrace{\Big(R(f_n) - R(f_{\mathcal{F}})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(f_{\mathcal{F}}) - R(f_{Bayes})\Big)}_{\text{approximation error}}$$



[von Luxburg and Schölkopf, 2008]

Underfitting *vs.* overfitting

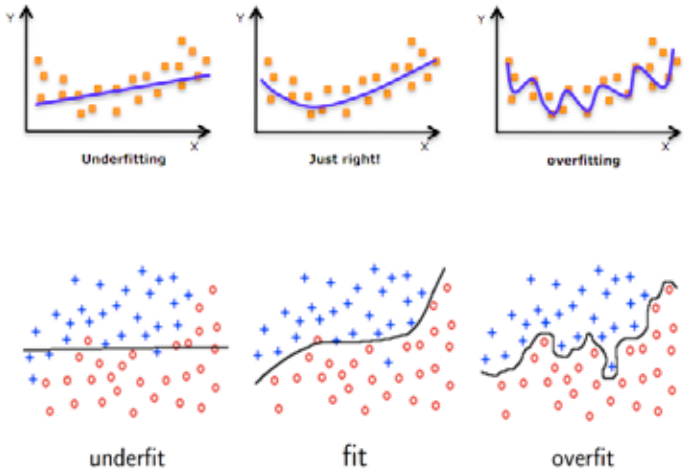


- ✓ small complexity of $\mathcal{F}$ ⇒ small estimation error, large approximation error (*underfitting*)
- ✓ large complexity of $\mathcal{F}$ ⇒ large estimation error, small approximation error (*overfitting*)

The best overall risk is achieved for "moderate" complexity

[von Luxburg and Schölkopf, 2008]

Model selection



Shattering

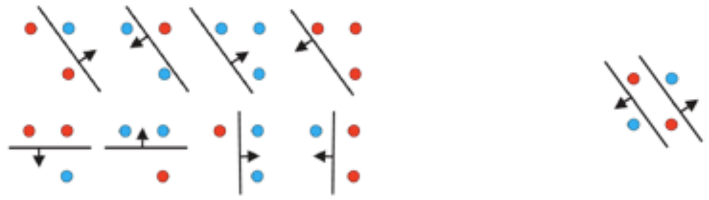
A set of n instances $X_1,\ldots, X_n$ from the input space $\mathcal{X}$ is said to be *shattered* by a function class $\mathcal{F}$ if all the 2^n labelings of them can be generated using functions from $\mathcal{F}$.

Example.

$\mathcal{F}$ = linear decision functions (straight lines) in the plane

(a) Any set of 3 non-collinear points shatters $\mathcal{F}$

(b) No set of 4 points can shatter $\mathcal{F}$



(a) (b)

The VC dimension

The *VC dimension* of a function class $\mathcal{F}$, denoted VC($\mathcal{F}$), is the largest integer h such that *there exists* a sample of size h which is shattered by $\mathcal{F}$.

If arbitrarily large samples can be shattered, then VC($\mathcal{F}$) = ∞.

Examples.

- ✓ $\mathcal{F}$ = linear decision functions in $\mathbf{R}^2$ ⇒ VC($\mathcal{F}$) = 3
- ✓ $\mathcal{F}$ = linear decision functions (hyperplanes) in $\mathbf{R}^n$ ⇒ VC($\mathcal{F}$) = $n + 1$
- ✓ $\mathcal{F}$ = multi-layer perceptrons with W weights ⇒ VC($\mathcal{F}$) = $O(W \log W)$
- ✓ $\mathcal{F}$ = nearest neighbor classifiers ⇒ VC($\mathcal{F}$) = ∞

VC dimension *vs.* number of parameters

> «In algebraic representation, the dimension of the set of curves depends upon the number of *parameters* whose values we may freely choose.
>
> We can therefore say that the number of freely determinable parameters of a set of curves by which a theory is represented is characteristic for the degree of falsifiability (or testability) of that theory.»

Karl Popper
The Logic of Scientific Discovery (1959)

Note. The VC dimension is in general not related to the number of free parameters of a model (e.g., $f_\alpha(x) = \text{sgn}(\sin(\alpha x))$: 1 parameter, VCdim = ∞).

Fundamental results

For all $f \in \mathcal{F}$, with probability at least $1 - \delta$, we have:

$$R(f) \le R_{\text{emp}}(f) + \sqrt{\frac{h(\log(2n/h)+1) - \log(\delta/4)}{n}}$$

where $h = \text{VC}(\mathcal{F})$, and n is the sample size.

With probability approaching 1, no matter what the unknown probability distribution, given more and more data, the expected error for the functions that ERM endorses at each stage eventually approaches the minimum value of expected error of the functions in $\mathcal{F}$ *if and only if* $\mathcal{F}$ has finite VC dimension.

Structural risk minimization

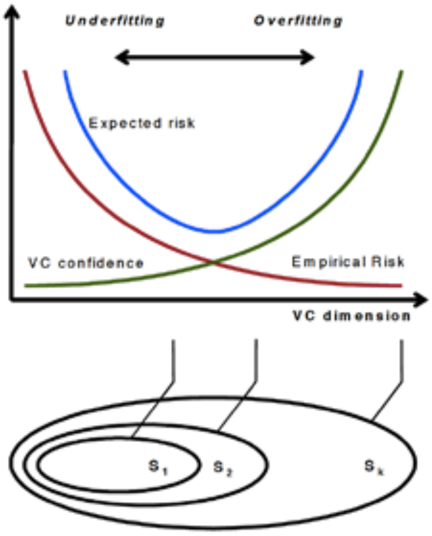
ERM takes only care of the *estimation* error (variance) but it is not concerned with the *approximation* error (bias).

The optimal model is found by striking a balance between the empirical risk and the capacity of the function class F (e.g., the VC dimension).

Basic idea of *Structural Risk Minimization* (SRM):

1. Construct a nested structure for family of function classes $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ with non-decreasing VC dimensions (VC($\mathcal{F}_1$) ≤ VC($\mathcal{F}_2$) ≤ …)
2. For each class $\mathcal{F}_i$, find the solution f_i that minimizes the empirical risk
3. Choose the function class $\mathcal{F}_i$, and the corresponding solution f_i that minimizes the risk bound (= empirical risk + VC confidence)

Structural risk minimization



Karl Popper as a precursor of SLT

> «Let me remark how amazing Popper's idea was. In the 1930's Popper suggested a general concept determining the generalization ability (in a very wide philosophical sense) that in the 1990's turned out to be one of the most crucial concepts for the analysis of consistency of the ERM inductive principles.»
>
> Vladimir Vapnik
> *The Nature of Statistical Learning Theory* (2000)

Further readings

G. Harman and S. Kulkarni, *Statistical learning theory as a framework for the philosophy of induction* (2008).

U. von Luxburg and B. Schölkopf. *Statistical learning theory: Models, concepts and results* (2008).

S. Kulkarni and G. Harman. *Statistical learning theory: A tutorial* (2011).