# Basic statistical learning - Part I

Scribes: Anjun Niu, Yewen Huang and An Chen

## 1 Statistical Learning Theory

Statistical learning is concerned with the problem: how can we learn a good model of an unknown distribution $P$ from sampled data?

To illustrate the challenge we face, we recall the classical problem of data regression. We have collected a set of measurements $y_1,\dots y_n \in Y$ at points $x_1,\dots,x_n \in X$, hence obtaining a dataset $D = \{(x_i, y_i)\}$. We are interest in computing a function $f : X \to Y$ such that $f(x)$ is a good predictor for $y \in Y$ a new data point $x \in X$.

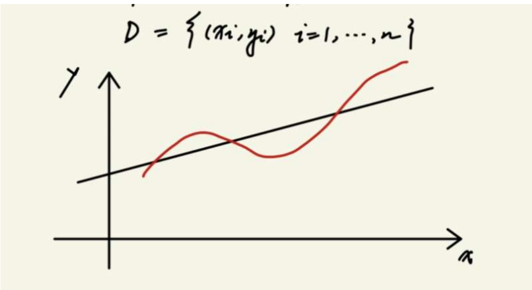

**Fig. 1** The data regression problem

In choosing the statistical model for the solution function model we face the so-called Bias-Variance dilemma. In the example, if we choose a linear model, then we require that all data can be described by a linear function, hence imposing a bias. If instead we choose as a model a polynomial of higher-degree, we can approximate better our data. However, little changes in the values $y_i$ (due, for instance, to measurement error) may cause large fluctuations in the model (cf. Fig. 1).

This observation leads to the following questions. What is the empirical risk telling us about the true risk? How can we ensure that our learning strategy converges?

## 1.1 Law of Large Numbers

To illustrate a version of the law of large numbers useful in our context, we set some notation first.

- Data: $D = \{(x_i, y_i) : \ i = 1, 2, \cdots, n, \ x_i \in X, \ y_i \in \{\pm 1\}\}$.
- Risk: $R_{L,P}(f) = \int_{X\times Y} L(x,y,f(x))dP(x,y)$.
- Empirical Risk: $R_{L,D}(f) = \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i))$.
- Loss:
$$L(x_i, f(x_i)) = \tfrac{1}{2}|f(x_i) - y_i| = \begin{cases} 0 & \text{if } f(x_i) = y_i \\ 1 & \text{if } f(x_i) \neq y_i. \end{cases}$$

We can interpret the quantities $\xi_i = L(x_i, f(x_i))$ as discrete random variables associated with Bernoulli trials ($\xi_i$ can take values 1 or 0).

We have the following classical estimate.

**Lemma 1 (Chernoff Bound (1952)).** *Let $\xi_1, \ldots, \xi_n$ be independent samples from a Bernoulli random variable $\xi$. Then*

$$P\{|\frac{1}{n}\sum_{i=1}^{n} \xi_i - E(\xi)| \geq \varepsilon\} \leq 2\exp(-2n\varepsilon^2).$$

That is, as the number of samples increases the difference between the empirical mean and the expectation of $\xi$ converges to 0 in probability.

The following is a generalization of the above lemma.

**Theorem 1 (Hoeffding Bound (1963)).** *Let $\xi_1, \ldots, \xi_n$ be independent samples from a bounded random variable $\xi$ with values in $[a,b]$. Let $Q_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i$. Then for any $\varepsilon > 0$,*

$$P(Q_n - E(\xi) \geq \varepsilon) \leq \exp(-\frac{2n\varepsilon^2}{(b-a)^2}),$$

$$P(E(\xi) - Q_n \geq \varepsilon) \leq \exp(-\frac{2n\varepsilon^2}{(b-a)^2}).$$

In order to prove Theorem 1, we recall the following classical inequality.

**Lemma 2 (Markov's inequality).** *Let $\xi$ be a non-negative r.v. with distribution $P$. For all $\lambda > 0$, we have*

$$P(\xi \geq \lambda E(\xi)) \leq \frac{1}{\lambda}.$$

*Proof.*

$$E(\xi) = \int_0^\infty \xi dP(\xi) \geq \int_{\lambda E(\xi)}^\infty \xi dP(\xi) \geq \lambda E(\xi) \int_{\lambda E(\xi)}^\infty dP(\xi) = \lambda E(\xi) P(\xi \geq \lambda E(\xi)).$$

**Proof of Theorem 1.** WLOG, let $E(\xi) = 0$ (if not, let $\bar{\xi} = \xi - E(\xi)$). To be able to apply Markov's inequality, we use the map $Q_n \rightarrow \exp(sQ_n)$ with $s > 0$. Hence, by Markov's inequality,

$$\begin{aligned} P(Q_n \geq \varepsilon) &= P(\exp(sQ_n) \geq \exp(s\varepsilon)) \\ &\leq e^{-s\varepsilon} E(e^{sQ_n}) \\ &= e^{-s\varepsilon} E(\exp(\frac{s}{n}\sum_{i=1}^n \xi_i)) \\ &= e^{-s\varepsilon} E(\Pi_{i=1}^n \exp \frac{s\xi_i}{n}) \\ &\leq e^{-s\varepsilon} \Pi_{i=1}^n E(\exp \frac{s\xi_i}{n}) \\ &\leq e^{-s\varepsilon} \exp(\frac{s^2(b-a)^2}{8n}). \end{aligned}$$

Note that the above inequality holds for any $s > 0$. The proof is completed by choosing $s = \frac{4n\varepsilon}{(b-a)^2}$. □

The result above shows that, *for fixed* $f$, we can $R_{D,L}(f) \xrightarrow{P} R_{P,L}(f)$, that is we achieve convergent in probability. In this case, the probability of a large deviation between $R_{D,L}(f)$ and $R_{P,L}(f)$ can be controlled (the larger the sample size the smaller the probability).

However, in practical learning problems, we are looking for a solution $f_{opt}$ that minimizes the risk while searching for a function $f_n$ that minimizes the empirical risk. That is, the function $f$ is not fixed and, thus, the result of Theorem 1 does not apply to this case. As illustrated in Fig. 2, the solution function $f_n$ that minimizes the empirical risk is gong to be different from the solution function $f_{opt}$ that minimizes the risk.

## 1.2 Consistency of empirical risk minimization

We want to identify conditions such that the function $f_n$ that minimizes $R_{D,L}(f)$ (here the index $n$ is the size of the training set $D$) is a good approximation of the function $f_{opt}$ that minimizes $R_{P,L}(f)$. This condition is associated with a notion of *consistency* that entails a restriction on the set of admissible functions $\mathscr{F}$ over which the empirical risk is minimized.

Since $f_{opt}$ is the solution of $\min_{f \in \mathscr{F}} R_{P,L}(f)$, then

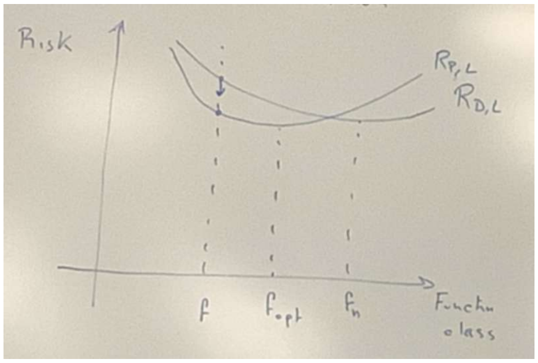

**Fig. 2** Intuitive illustration of the convergence of the empirical risk as compared to the actual risk.

$$R_{P,L}(f) - R_{P,L}(f_{opt}) \geq 0, \; \forall f \in \mathscr{F}.$$

Similarly, since $f_n$ the solution of $\min_{f \in \mathscr{F}} R_{D,L}(f)$, we have that

$$R_{D,L}(f) - R_{D,L}(f_n) \geq 0, \; \forall f \in \mathscr{F}.$$

It follows that

$$R_{P,L}(f_n) - R_{P,L}(f_{opt}) \geq 0 \quad \text{and} \quad R_{D,L}(f_{opt}) - R_{D,L}(f_n) \geq 0.$$

Thus

$$\begin{aligned} 0 &\leq (R_{P,L}(f_n) - R_{P,L}(f_{opt})) + (R_{D,L}(f_{opt}) - R_{D,L}(f_n)) \\ &= (R_{P,L}(f_n) - R_{D,L}(f_n)) + (R_{D,L}(f_{opt}) - R_{P,L}(f_{opt})) \\ &\leq \sup_{f \in \mathscr{F}} (R_{P,L}(f) - R_{D,L}(f)) + (R_{D,L}(f_{opt}) - R_{P,L}(f_{opt})) \end{aligned}$$

By law of large numbers, $R_{D,L}(f_{opt}) \xrightarrow{P} R_{P,L}(f_{opt})$ as $n \to \infty$. It follows that, if

$$\sup_{f \in \mathscr{F}} (R_{P,L}(f) - R_{D,L}(f)) \xrightarrow{P} 0, \; \text{as } n \to \infty, \tag{1}$$

then $R_{P,L}(f_n) \xrightarrow{P} R_{P,L}(f_{opt})$ and $R_{D,L}(f_{opt}) \xrightarrow{P} R_{D,L}(f_n)$, that is, we obtain *consistency of the empirical risk minimization of the class of functions $\mathscr{F}$*.

The argument above shows that one-sided uniform convergence (1) gives a sufficient condition for consistency. In fact, this condition is also necessary.

**Theorem 2.** *One-sided uniform converges in probability*

$$\lim_{n \to \infty} P[\sup_{f \in \mathscr{F}} (R_{P,L}(f) - R_{D,L}(f)) > \varepsilon] = 0,$$

*for all $\varepsilon > 0$, is a necessary and sufficient condition for non-trivial consistency of empirical risk minimization (ERM).*

The theorem shows that consistency and, hence, learning, depends critically on the selection of the set of admissible functions $\mathscr{F}$ over which the empirical risk is minimized.

Below we take a closer look at the implications of Theorem 2 and derive conditions on the class $\mathscr{F}$ guaranteeing consistency of the empirical risk minimization. There are two main ideas we explore: union bound + symmetrization

To introduce union bounds, let us start by examining the simple case where $\mathscr{F} = \{f_1, f_2\}$. For $i = 1,2$, let

$$C_\varepsilon^i = \{(x_1,y_1),\cdots,(x_n,y_n) : R_{P,L}(f_i) - R_{D,L}(f_i) > \varepsilon\}.$$

It follows that

$$\begin{aligned} P(\sup_{f\in\mathscr{F}}(R_{P,L}(f) - R_{D,L}(f)) > \varepsilon) &= P(C_\varepsilon^1 \cap C_\varepsilon^2) \\ &= P(C_\varepsilon^1) + P(C_\varepsilon^2) - P(C_\varepsilon^1 \cap C_\varepsilon^2) \\ &\leq P(C_\varepsilon^1) + P(C_\varepsilon^2). \end{aligned}$$

with equality holding iff events are disjoint. More generally, if $\mathscr{F} = \{f_1, f_2, \cdots, f_m\}$, then

$$P(\sup_{f\in\mathscr{F}}(R_{P,L}(f) - R_{D,L}(f)) > \varepsilon) \leq \sum_{i=1}^{m} P(C_\varepsilon^i).$$

The last inequality is called the *union bound* and it shows how to manage the situation when $\mathscr{F}$ is a finite set. To deal with infinite case, we use the following symmetrization result due to [Vapnik, Chervonenkis, 1979].

**Lemma 3 (Symmetrization).** *For $m\varepsilon^2 > 2$ we have*

$$P[\sup_{f\in\mathscr{F}}(R_{P,L}(f) - R_{D,L}(f)) > \varepsilon] \leq 2P[\sup_{f\in\mathscr{F}}(R_{D,L}(f) - R'_{D,L}(f)) > \varepsilon/2],$$

*where the first P refers to the distribution of i.i.d. samples of size n and the second P refers to the distribution of i.i.d. samples of size $2n$; in the latter case, $R_{D,L}$ measures the loss on the first half of the samples and $R'_{D,L}$ measures the loss on the second half.*

Lemma 3 shows that class $\mathscr{F}$ is effectively finite. The empirical risk defines functions by their values over $m$ points, or $2m$, as in the right hand side of the lemma. Since at each point a function can only take 2 possible values, there are at most $2^{2m}$ possible elements in $\mathscr{F}$ as defined by their value at $2m$ points).

For a $2m$ sample set $D_{2m} = \{(x_1,y_1),...,(x_{2m},y_{2m})\}$, we define $N(\mathscr{F}, D_{2m}) =$ to be the cardinality of $\mathscr{F}$ when restricted to $x_1,...,x_{2m}$. That is, it counts the number of functions in $\mathscr{F}$ that can be distinguished by their values on $x_1,...,x_{2m}$. The function $N(\mathscr{F}, 2m)$ counts the maximum number of functions that can be distinguished (over all possible choices of $2m$ samples) by their values $x_1,...,x_{2m}$. $N(\mathscr{F},2m)$ is called the *shattering coefficient* of $\mathscr{F}$ and measures the number of ways the class $\mathscr{F}$ can separate the patterns into 2 classes. If $N(\mathscr{F},2m) = 2^{2m}$, then all possible separation can be implemented by function in this class and we say that $\mathscr{F}$ shatters $2m$ points. Note that this means that there exists a set of $2m$ patterns that can be separated in all possible way but it does not necessarily apply to all sets of $2m$ patterns.

Using Lemma 3, the law of large numbers and the union bound, Vapnik and Chervonenkis derived the following estimate.

$$P[\sup_{f\in\mathscr{F}}(R_{P,L}(f)-R_{D,L}(f)) > \varepsilon] \leq 4E[N(\mathscr{F},D_{2m})]\exp(-\frac{m\varepsilon^2}{8})$$

$$= 4\exp(\ln E[N(\mathscr{F},D_{2m})]-\frac{m\varepsilon^2}{8}), \qquad (2)$$

where the term $\ln E[N(\mathscr{F},D_{2m})]$ is known as the *annealed entropy*.

This shows that, provided $E[N(F,D_{2m})]$ does not grow exponentially with respect to$m$, then one can derive a non-trivial bound about the test error.

From (2), we can derive a bound on $R_{P,L}(f)$. For that, set the RHS of the inequality equal to $\delta > 0$, then solve for $\varepsilon$. We obtain that, with probability at least $1-\delta$, we have

$$R_{P,L}(f) \leq R_{D,L}(f) + \sqrt{\frac{8}{m}(\ln E[n(F,D_{2m})] + \ln\frac{4}{\delta})}, \qquad (3)$$

where the second term on the RHS of the inequality is called the *confidence* or *capacity term*

We remark that the bound is independent of $f$ and hods, in particular, for the function $f_m$ minimizing the empirical risk. The capacity term is a property of the function class $F$ and not of the individual function $f$. It follows that the bound cannot be minimized over a specific $f$. Instead, we can introduce a structure on $\mathscr{F}$ and minimize over the elements of the structure leading to Structure Risk Minimization.

The capacity term expressed in terms of the annealed entropy. $InE[N(\mathscr{F},D_{2m})]$ is impractical to evaluate. As a result, alternative bounds have been proposed in the literature. We describe below how to derive the notion of VC dimension.

For a single data point $(x,y)$, $f$ causes a loss $L(y,f(x)) = \frac{1}{2}|f(x)-y|$. For a larger sample $D_m = (x_1,y_1),...,(x_m,y_m)$, the loss vector is $\xi_f = (L(y_1,f(x_1)),...,L(y_m,f(x_m)))$ whose cardinality is $N(\mathscr{F},D_m)$. The *VC entropy* is defined as

$$H_{\mathscr{F}}(m) = E[\ln N(\mathscr{F},D_m)],$$

where $E$ is taken over the random generation of $m$ samples from $P$. One can show that the condition

$$\lim_{m\to\infty}\frac{1}{m}H_{\mathscr{F}}(m) = 0$$

is equivalent to

$$\lim_{n\to\infty}P[\sup_{f\in\mathscr{F}}(R_{L,P}(f)-R_{L,D}(f)) > \varepsilon] = 0 \quad \forall\varepsilon > 0$$

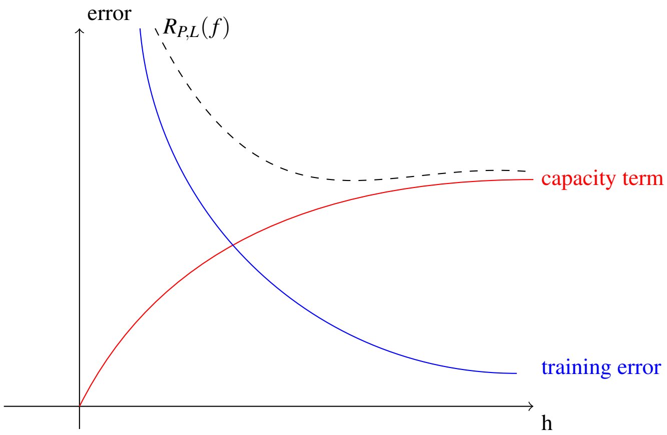

**Fig. 3** Graphical illustration of risk minimization.

which implies the consistency of empirical risk minimization Theorem 2.

By exchanging the expectation and the logarithm in the definition of $H_{\mathscr{F}}(m)$ we have

$$H_{\mathscr{F}}^{ann} = \ln E[N(\mathscr{F}, D_m)] = \text{annealed entropy}.$$

Since ln is concave, then

$$H_{\mathscr{F}}(m) \leq H_{\mathscr{F}}^{ann}(m).$$

One can show that the condition

$$\lim_{m\to\infty} \frac{1}{m} H_{\mathscr{F}}^{ann}(m) = 0$$

is necessary and sufficient to have

$$P[\sup_{f\in\mathscr{F}} (R_{P,L}(f) - R_{D,L}(f)) > \varepsilon] \leq 4\exp((\frac{1}{m} H_{\mathscr{F}}^{ann}(2m) - \varepsilon^2)\cdot m).$$

Next, we define the *growth function*

$$G_{\mathscr{F}}(m) = \max_{D_m \in \mathscr{X}\times(\pm 1)} \ln N(\mathscr{F}, D_m) = \ln N(\mathscr{F}, m).$$

The convergence

$$\lim_{n\to\infty} \frac{1}{m} G_{\mathscr{F}}(m) = 0 \tag{4}$$

is necessary and sufficient for exponentially fast convergence of risk for all underlying distributions $P$.

If $\mathscr{F}$ is a very rich class so that, for any sample of size $m$, the points can be sheltered, then $G_{\mathscr{F}}(m) = m\ln(a)$. In this case, (4) does not hold and learning is not successful. In the other case, there exists a maximal $m$ for which (4) holds. This number is called the *VC dimension* and is denoted with $h$. Hence, the VC dimension is the maximal number of points that can be shattered by a function in $\mathscr{F}$. For $m \leq h$, the growth function $G_F(m)$increases linearly with the sample size. If $m > h$, it grows only logarithmically and this is the situation where learning can succeed. We have:

$$G_{\mathscr{F}}(m) \leq h(\ln m/h + 1).$$

In fact, we have the following succession of capacity concepts and corresponding inequalities

$$H_{\mathscr{F}}(m) \leq H_{\mathscr{F}}^{ann} \leq G_{\mathscr{F}}(m) \leq h(\ln m/h + 1).$$

Using these observations, we derive the following estimate of the risk in terms of the VC dimension $h$:

$$R_{P,L}(f) \leq R_{D,L}(f) + \sqrt{\frac{1}{m} h(\ln \frac{2m}{h} + 1) + \ln \frac{4}{\delta}},$$

which holds with probability$< 1 - \delta$. This is illustrated in Fig. 3.

## *1.3 Example of VC dimension*

Let $x_1, x_2, x_3 \in \mathbf{R}^2$ be not collinear points and consider the problem of classifying binary patters using functions in the class $\mathscr{F} = \{\text{hyperplanes in}\mathbf{R}^2\}$. There are $2^3 = 8$ ways to assign the 3 points to 2 classes and all 8 assignments can be realized with separating lines. That is, we can always find $f \in \mathscr{F}$ with $f(x_i) = y_i \ \forall i$ (see Fig. 4). This shows that the VC dimension of $\mathscr{F}$ is $h \geq 3$. Since we can never shatter four points in $\mathbb{R}^2$ using functions in $\mathscr{F}$, then $h = 3$.

Using a similar argument, one can show that the VC dimension of the class $\mathscr{F}$ of hyperplanes in $\mathbf{R}^N$ has VC dimension $h = N + 1$.

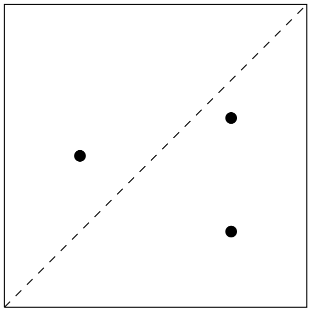

**Fig. 4** Three points in $\mathbb{R}^3$ can always be separated in two groups by a line provided they are not collinear.

Finally, we examine the VC dimension associated with the SVM problem. While in this case the dimnsion of the featire space can be even infinite, however the SVM problem invilves hyperplanes with margins which has the consequence of reducing the space capacity.

**Theorem 3 (Vapnik, 1979).** *Consider hyperlane* $< w,x >= 0$ *where* $w$ *is normalized such that they are in canonical form with respect to* $X\{x_1,\ldots,x_r\}$*, i.e.,* $\min_{(1,\ldots,r)} | < w,x_i > | = 1$*. The set of decision functions* $f_w =< w,x >$ *define on* $X$ *and satisfying the constraint* $\| w \|< \Lambda$ *has VC dimension satisfying* $h \leq \mathbf{R}^2\Lambda^2$*, where* $\mathbf{R}$ *is the radius of smallest ball centered at origins and containing* $X$*.*

Therefore, according to the theorem, one can control the VC dimension irrespective of the dimension of the space by controlling the quantity $\| w \|$.