A short introduction to statistical learning theory

MATH 637

1. PAC Learning and ERM

Standard framework for supervised learning: hypothesis space, loss and risk. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the output space, supervised learning tries to learn a function that maps an input to an output based on example input-output pairs.

- Rigorously, given a sequence of label data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ sampled (independently and identically) from an unknown distribution $P_{X,Y}$, a learning algorithm seeks a function $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space and h belong a set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ (which we refer to as the *hypothesis space*).
- To measure how well a function fits the training data, a *loss function* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ is defined. For training example (x_i, y_i) and a hypothesis h, the loss of predicting the value $h(x_i)$ is $L(h(x_i), y_i)$.
- This pre-defined loss function induces a *risk function* on $\mathcal{H}$, defined as

  $$R(h) := \mathbb{E}_{(X,Y)\sim P}[L(h(X), Y)].$$

  For a hypothesis h, $R(h)$ is the expected loss incurred per sample when h is used to make the prediction.
- The "optimal hypothesis" h^*, whose performance we wish to replicate, is the minimizer over $\mathcal{H}$ of the risk function $R(h)$.

Empirical risk minimizer (ERM)

As mentioned above, an algorithm takes as input a finite sequence of training samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and outputs a function from $\mathcal{X} \to \mathcal{Y}$. The most standard algorithm is the *empirical risk minimizer* (ERM), which outputs

$$\hat{h}_n = \min_{h\in\mathcal{H}} R_n(h)$$

where

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i), y_i)$$

is the *empirical risk*.

The main idea of ERM is that since

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, h(x_i)) \approx \mathbb{E}_{(X,Y)\sim P}[L(Y, h(X))] = R(h),$$

minimizing $R_n(h)$ will have a similar effect as minimizing $R(h)$. However, we observe (through simulation) that ERM doesn't always work as expected, and we need a good theory to better understand why and how it fails. To establish that rigorously, we need to quantify explicitly the event for which $R_n(h)$ is close to $R(h)$ for each hypothesis h.

PAC Learning. The probably approximately correct (PAC) learning model typically states as follows: we say that $\hat{h}_n$ is ϵ-accurate with probability $1-\delta$ if

$$P\left[R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) > \epsilon\right] < \delta.$$

In other words, we have $R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \leq \epsilon$ with probability at least $(1-\delta)$.

2. *Hoeffding's inequality*

Concentration inequalities provide bounds on how a random variable deviates from some value (typically, its expected value). In this section, we sketch the main steps to derive a class of concentration inequalities for bounded random variables.

The underlying idea is to upper-bound a tail probability $P[X \geq t]$ by controlling the moments of the random variable X.

Theorem 1 (Markov inequality)**.** *For any nonnegative random variable X and $\epsilon > 0$,*

$$P[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

Theorem 2. *For any random variable X, $\epsilon > 0$ and $t > 0$*

$$P[X \geq \epsilon] \leq \frac{\mathbb{E}[e^{tX}]}{e^{t\epsilon}}.$$

Theorem 3. *If random variable X has mean zero and is bounded in $[a,b]$, then for any $s > 0$,*

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

Theorem 4 (Hoeffding's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d copies of a random variable $X \in [a,b]$, and $\epsilon > 0$,*

$$P\left[\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right).$$

Proof. We have

$$\begin{aligned} & P\left[\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X] \geq \epsilon\right] \\ &= P\left[(X_1 + X_2 + \ldots + X_n) - E[X_1 + X_2 + \ldots + X_n] \geq n\epsilon\right] \\ &\leq \exp(-tn\epsilon)\mathbb{E}[e^{t[(X_1+X_2+\ldots+X_n) - E[X_1+X_2+\ldots+X_n]]}] \\ &= \exp(-tn\epsilon)\prod_{i=1}^{n}\mathbb{E}[e^{t[X_i - E[X_i]]}] \\ &\leq \exp\left(-tn\epsilon + n\frac{t^2(b-a)^2}{2}\right) \end{aligned}$$

Note: We can apply Theorem 3 for $X_i - EX_i$ in the bounds above because

$$E[X_i - EX_i] = 0 \quad \text{and} \quad -(b-a) \leq X_i - EX_i \leq (b-a).$$

The quadratic expression (in t) attains maximum value at

$$t = \frac{\epsilon}{(b-a)^2}.$$

Replacing this value of t in the inequality, we deduce that

$$P\left[\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X] \geq \epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right).$$

Using a similar argument, we also have

$$P\left[\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X] \leq -\epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right).$$

The combination of those two estimates completes the proof. □

3. *Generalization bound for finite hypothesis space and bounded loss.*

Assume that

- the loss function L is bounded, that is

$$0 \leq L(y, y') \leq c \quad \forall y, y' \in \mathcal{Y}$$

- the hypothesis space is a finite set, that is

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}.$$

Using the Hoeffding's inequality, for any $h \in \mathcal{H}$ and $\epsilon > 0$ we have

$$P[|R_n(h) - R(h)| \geq \epsilon] \leq 2\exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

Thus

$$P[\exists h \in \mathcal{H} : |R_n(h) - R(h)| \geq \epsilon] \leq 2|\mathcal{H}|\exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

This means that, with probability at least

$$1 - 2|\mathcal{H}|\exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

we have

$$R(\hat{h}_n) - R(h^*) = [R(\hat{h}_n) - R_n(\hat{h}_n)] + [R_n(\hat{h}_n) - R_n(h^*)] + [R_n(h^*) - R(h^*)] \leq 2\epsilon$$

(note that the second term is non-positive by the definition of the ERM).

Thus, for any $\delta > 0$ and $\epsilon > 0$, by choosing

$$n = \frac{2c^2}{\epsilon^2}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

then $\hat{h}_n$ is ϵ-accurate with probability $1-\delta$, i.e.

$$P\left[R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) > 2\epsilon\right] < \delta.$$

Corollary. If we quantify the error in terms of number of samples, then

$$R(\hat{h}_n) \leq R(h^*) + \frac{c}{\sqrt{n}}\sqrt{8\log\left(\frac{2}{\delta}\right) + 8\log(|\mathcal{H}|)}.$$

4. *Generalization bound using covering number*

We know that for finite hypothesis space and bounded loss, if we quantify the error in terms of the number of samples, then

$$|R_n(h) - R(h)| \leq \frac{c}{\sqrt{n}}\sqrt{2\log\left(\frac{2}{\delta}\right) + 2\log(|\mathcal{H}|)}, \forall h \in \mathcal{H}$$

with probability at least $1-\delta$.

What about infinite hypothesis classes?

Assumption. In this note, we assume that $\mathcal{H}$ is a metric space with distance d defined on it. For $\epsilon > 0$, we denote by $\mathcal{N}(\epsilon,\mathcal{H},d)$ the *covering number* of $(\mathcal{H},d)$; that is, $\mathcal{N}(\epsilon,\mathcal{H},d)$ is the minimal number of balls of radius ϵ needed to cover $\mathcal{H}$. We denote by $\mathcal{H}_\epsilon$ a finite subset of $\mathcal{H}$ such that $\mathcal{H}$ is contained in the union of balls of radius ϵ and $|\mathcal{H}_\epsilon| = \mathcal{N}(\epsilon,\mathcal{H},d)$.

Note: If $\mathcal{H}$ is a $dk-$dimensional manifold/algebraic surface, then we now that

$$\mathcal{N}(\epsilon,\mathcal{H},d) = \mathcal{O}\left(\epsilon^{-k}\right)$$

Assume further that the loss function L satisfies:

$$|L(h(x),y) - L(h'(x),y)| \leq Cd(h,h') \quad \forall, x\in\mathcal{X}; y\in\mathcal{Y}; h,h'\in\mathcal{H}$$

Generalization bound using covering number.

We first note that if

$$n = \frac{8c^2}{\epsilon^2}\log\left(\frac{2|\mathcal{H}_\epsilon|}{\delta}\right)$$

then the event

$$|R_n(h) - R(h)| \leq \epsilon, \forall h \in \mathcal{H}_\epsilon$$

happens with probability at least $1-\delta$.

Under this event, consider any $h \in \mathcal{H}$, then there exists $h_0 \in \mathcal{H}_\epsilon$ such that $d(h,h_0) \leq \epsilon$. This means

$$|R_n(h) - R_n(h_0)| \leq Cd(h,h_0)$$

and

$$|R(h) - R(h_0)| \leq Cd(h,h_0).$$

This implies that

$$|R_n(h) - R(h)| \leq (2C+1)\epsilon \quad \forall h \in \mathcal{H}.$$

We conclude that for all $\epsilon > 0$, $\delta > 0$, if

$$n = \frac{8c^2}{\epsilon^2} \log\left(\frac{2\mathcal{N}(\epsilon, \mathcal{H}, d)}{\delta}\right)$$

then

$$|R_n(h) - R(h)| \leq (2C+1)\epsilon \quad \forall h \in \mathcal{H}.$$

with probability at least $1 - \delta$.