Topic 10: STATISTICAL LEARNING THEORY

STAT 37710/CMSC 25400 Machine Learning
Risi Kondor, The University of Chicago

Back to Supervised Learning

- **Training set:** $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$
- **Assumption:** each (x, y) is chosen IID from some distribution p on $\mathcal{X} \times \mathcal{Y}$
- **Hypothesis:** a mapping $f : x \mapsto y$ chosen from some hypothesis space $\mathcal{F}$
- **Loss function:** $\ell_{\text{true}}(\widehat{y}, y) = \mathbb{I}(\widehat{y} \neq y)$ $(0/1$ loss)
- **Goal:** find an $\widehat{f} \in \mathcal{F}$ with low true error

$$\mathcal{E}_{\text{true}}[\widehat{f}] = \mathbb{E}_{(x,y)\sim p}\, \ell(\widehat{f}(x), y).$$

Frequentist (discriminative) approach: just focus on finding a good $\widehat{f}$. Don't worry about learning p.

Regularized Risk Minimization (RRM)

Finds $\widehat{f} \in \mathcal{F}$, which minimizes the regularized risk

$$\mathcal{E}_S^{\text{reg}}[f] = \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i)}_{\text{training error}} + \underbrace{\lambda\,\Omega[f]}_{\text{regularizer}}$$

But how well will $\widehat{f}$ do on future examples? What is its true error???

$\rightarrow$ Statistical Learning Theory

Empirical error vs. true error

- What we can measure (and what we optimize for) is the empirical error on the training set

$$\mathcal{E}_S[f] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i) \neq y_i).$$

- What we want to bound is the true error

$$\mathcal{E}_{\text{true}}[\widehat{f}] = \mathbb{E}_{(x,y)\sim p}\,\ell(\widehat{f}(x), y).$$

Question: Does a low $\mathcal{E}_S$ imply a low $\mathcal{E}_{\text{true}}$? Yes, provided we are not overfitting.

Probably Approximately Correct bounds

Can we show (without knowing p) that for some small ϵ

$$\mathcal{E}_{\text{true}}(\widehat{f}) \leq \mathcal{E}_S(\widehat{f}) + \epsilon? \tag{1}$$

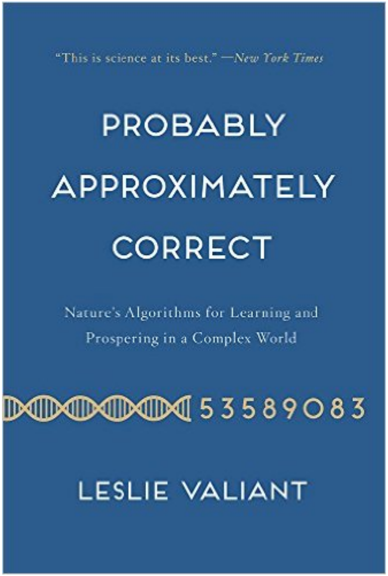
No, because a really misleading training set can always mess us up.

Let's say that a training set S is **evil** if for the $\widehat{f}$ that our algorithm returns, (1) is violated. PAC style bounds show that $\mathbb{P}[S \text{ is evil}] < \delta$, i.e.,

$$\mathbb{P}\Big[\, \mathcal{E}_{\text{true}}(\widehat{f}) > \mathcal{E}_S(\widehat{f}) + \epsilon \,\Big] < \delta$$

for some small probability δ (over draws of S).

[Valiant, 1984]


"This is science at its best." —New York Times
PROBABLY
APPROXIMATELY
CORRECT
Nature's Algorithms for Learning and
Prospering in a Complex World
53589083
LESLIE VALIANT

Hoeffding bound

For any given f with $\mathcal{E}_{\text{true}}[f] = \pi$, whether or not f makes a mistake on a random $(x, y) \sim p$ is just a $\text{Bernoulli}(\pi)$ random variable.

Hoeffding bound: If $X_1, X_2, \ldots, X_m \sim^{\text{IID}} \text{Bernoulli}(\pi)$, then

$$\mathbb{P}\left[\,\tfrac{1}{m}\textstyle\sum_{i=1}^{m} X_i < \pi - \epsilon\,\right] \leq e^{-2m\epsilon^2}.$$

Therefore, with probability $1 - \delta$ we can guarantee that the difference in error is less than

$$\epsilon = \sqrt{\frac{\log(1/\delta)}{2m}}.$$

This is how hold-out sets work.

A false argument

Since $X_i = \mathbb{I}(\widehat{f}(x_i) \neq y_i)$ are IID Bernoulli$(\mathcal{E}_{\text{true}}(\widehat{f}))$ random variables,

$$\mathbb{P}[S \text{ is evil}] = \mathbb{P}\left[\, \mathcal{E}_S(\widehat{f}) < \mathcal{E}_{\text{true}}(\widehat{f}) - \epsilon \,\right] \leq e^{-2m\epsilon^2}.$$

Question: What is the problem here?

The hypothesis $\widehat{f}$ also depends on S, so given $\widehat{f}$, the random variables $X_1, X_2, \ldots, X_m$ are not distributed according to the same distribution as a general $X = \mathbb{I}(\widehat{f}(x) \neq y)$, and give an overoptimistic estimate of $\mathcal{E}_{\text{true}}[\widehat{f}]$.

In fact, the $\widehat{f}$ chosen by ERM/RRM tends to be one for which $\mathcal{E}_{\text{true}} - \mathcal{E}_S$ is particularly high. $\to$ This is not just a theoretical difficulty.

In practice, can always use a holdout set. $\to$ Honest answer, but doesn't tell us anything about why ERM actualy works.

The union bound

Idea of **Uniform convergence**: put a bound on

$$\mathbb{P}\,[\,\exists f \in \mathcal{F} \quad \mathcal{E}_S(f) < \mathcal{E}_{\text{true}}(f) - \epsilon\,] \geq \mathbb{P}[S \text{ is evil}]$$

The event on the left does not depend on $\widehat{f}$, so now the (x_i, y_i)'s really are IID.

If $\mathcal{F}$ is a finite set of cardinality C, we have the **union bound**:

$$\mathbb{P}\,[\,\exists f \in \mathcal{F} \quad \mathcal{E}_S(f) < \mathcal{E}_{\text{true}}(f) - \epsilon\,] \leq C\,e^{-2m\epsilon^2},$$

giving

$$\epsilon = \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2m}}.$$

This is a huge overkill and only works for finite hypothesis spaces. (There are lots of $f \in \mathcal{F}$, but they are not all that different in behavior.)


Vapnik–Chervonenkis theory

How do we quantify just how prone $\mathcal{F}$ is to overfitting?

Key idea

Take an independent **ghost sample** S' of size m from p (bit like a virtual hold-out set) and prove

$$\mathcal{E}_S[\widehat{f}] \text{ is low} \implies \mathcal{E}_{S'}[\widehat{f}] \text{ is low} \implies \mathcal{E}_{\text{true}}[\widehat{f}] \text{ is low.}$$

This reduces to computing the union bound wrt $\mathcal{F}{\downarrow_{S \cup S'}}$.

For simplicity, in the following slides assume the simplest case of $\mathcal{E}_S[\widehat{f}] = 0$

Idea 1: Symmetrization

Any $f \in \mathcal{F}$ splits $\overline{S} = S \cup S'$ into two sets:

1. the mistake points $E_f = \{ (x, y) \in \overline{S} \mid f(x) \neq y \}$
2. the correct points $E'_f = \{ (x, y) \in \overline{S} \mid f(x) = y \}$.

We say that f is **bad** if $|E_f| \geq k := \lfloor m\epsilon/2 \rfloor$, but all the mistakes are in S'.

- Given $x_1, \ldots, x_{2m}$ and an f with $|E_f| \geq k$, what is the probability that it is **bad**?

$$p \leq \binom{m}{k} \Big/ \binom{2m}{k} = \frac{m(m-1)\ldots(m-k+1)}{2m(2m-1)\ldots(2m-k+1)} \leq 2^{-k}.$$

- Now what is the probability that there is some $f \in \mathcal{F}$ that is bad? By the union bound:

$$p \leq 2^{-k} \, |\mathcal{F}\downarrow_{\overline{S}}|,$$

where $|\mathcal{F}\downarrow_{\overline{S}}|$ is the number of ways that $\mathcal{F}$ can carve up $\overline{S}$ into $E_f \cup E'_f$.

Idea 2: Vapnik–Chervonenkis dim

Definition

We say that a set $V \subseteq \mathcal{X}$ is shattered by $\mathcal{F}$ if $|\mathcal{F}{\downarrow_V}| = 2^{|V|}$. The VC–dimension d of $\mathcal{F}$ is the cardinality of the largest $V \subseteq \mathcal{X}$ that is shattered by $\mathcal{F}$.

Examples:

For linear classifiers in $\mathbb{R}^n$, $d = n + 1$.

For axis-aligned rectangles in $\mathbb{R}^n$, $d = 2n$.

Lemma (Sauer–Shelah)

If the VC–dimension of $\mathcal{F}$ is d, then for any $V \subseteq \mathcal{X}$ of cardinality m,

$$|\mathcal{F}{\downarrow_V}| \leq \left(\frac{em}{d}\right)^d.$$

Idea 3: Chernoff bound

Theorem

If $X_1, X_2, \ldots, X_m$ are independently distributed binary random variables with $\mathbb{P}(X_i = 1) = \theta$ and $\mathbb{P}(X_i = 0) = 1 - \theta$, then

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} X_i < (1-\gamma)\,\theta\right] < e^{-m\theta\gamma^2/2}.$$

Corollary

For any f, and any IID sample S' of size $m \geq 8/\mathcal{E}_{\mathrm{true}}(f)$,

$$\mathbb{P}\left[\,\mathcal{E}_{S'}(f) < \mathcal{E}_{\mathrm{true}}(f)/2\,\right] < 0.5.$$

Putting it all together

$\mathbb{P}\big[\,\mathcal{E}_{\text{true}}(\widehat{f}) > \epsilon\,\big]$

$\leq 2\,\mathbb{P}\,[\,\mathcal{E}_{S'}(f) > \epsilon/2\,]$ (Chernoff)

$\leq 2 \cdot 2^{-\lfloor m\epsilon/2 \rfloor}\,|\mathcal{F}\!\downarrow_{\overline{S}}|$ (symmetrization and using $\mathcal{E}_S(\widehat{f}) = 0$)

$\leq 2 \cdot 2^{-\lfloor m\epsilon/2 \rfloor} \left(\frac{2em}{d}\right)^d$ (Sauer-Shelah)

$< \delta$ (this is what we require)

In the $\mathcal{E}_S(\widehat{f}) > 0$ case the analysis is only a shade more involved.

A general VC–bound

Theorem

If $\mathcal{F}$ is a hypothesis class over $\mathcal{X}$ of VC–dimension d, then

$$\mathbb{P}\left[\mathcal{E}_{\mathrm{true}}(\widehat{f}) > \mathcal{E}_S(\widehat{f}) + \sqrt{\frac{d(\log\frac{2m}{d}+1)+\log(4/\delta)}{m}}\right] \leq \delta.$$

Margin–based VC bound

If $\mathcal{F}_\gamma$ is the space of hyperplanes with margin $\geq \gamma$ in $\mathbb{R}^n$ and $\widehat{f} \in \mathcal{F}$, then

$$\mathbb{P}\left[\mathcal{E}_{\text{true}}(\widehat{f}) > \mathcal{E}_S(\widehat{f}) + \sqrt{\frac{\frac{1}{\gamma^2}(\log(2m\gamma^2)+1)+\log(4/\delta)}{m}}\right] \leq \delta.$$

Rademacher averages

Effective hypothesis space

First, put (x, y) together into a single variable z, and for $f \in \mathcal{F}$ define

$$f'(z) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{if } f(x) \neq y \end{cases}.$$

Notice that

$$\mathbb{P}[\mathcal{E}_{\text{true}}(\widehat{h}) > \mathcal{E}_S(\widehat{h}) + \epsilon] \leq \delta \quad \Longleftarrow \quad \mathbb{P}\Big[\sup_{f' \in \mathcal{F}'} [\mathbb{E} f'(z) - \mathbb{E}_S f'(z)] > \epsilon\Big] \leq \delta,$$

where $\mathcal{F}' = \{ f' \mid f \in \mathcal{F} \}$ is the effective hypothesis class. $\rightarrow$ For simplicity, in the following work with $\mathcal{F}'$, but drop the dashes.

Rademacher average

The **Rademacher average** of $\mathcal{F}$ (w.r.t. the unknown distribution p) is

$$R_m(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i)\right],$$

where $z_1, \ldots, z_m \sim p$ and $\sigma_1, \ldots, \sigma_m$ are independent Rademacher random variables (i.e., $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$).

The **empirical Rademacher average** given $S = \{z_1, \ldots, z_m\}$ is

$$\widehat{R}_m(\mathcal{F}) = \mathbb{E}_{\sigma_1, \ldots, \sigma_m}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i)\right].$$

SUMMARY

- Instead of trying to prove $\mathcal{E}_{\text{true}}[\widehat{f}] \leq \epsilon$, prove $\mathbb{P}[\mathcal{E}_{\text{true}}(\widehat{f}) > \epsilon] < \delta$.
- Given $\widehat{f}$, the training set is no longer IID from p $\rightarrow$ union bound.
- Instead of $|\mathcal{F}|$, characterize the complexity of $\mathcal{F}$ by its behavior on a finite sample $\rightarrow$ VC-dimension.
- $\mathcal{E}_{S'}[\widehat{f}]$ is concentrated around its mean $\rightarrow$ Chernoff bound.
- The probability that all the errors in $S \cup \overline{S}$ will be in $\overline{S}$ and none in S is small $\rightarrow$ symmetrization.
- VC-bounds are outmoded. Nowadays people use Rademacher averages and stronger concentration results.
- For practical purposes the bounds are way too loose. Things can be bad, but they are usually not as bad as they could be.

FURTHER READING

- L Valiant: A Theory of the Learnable (1984)
- V. Vapnik: Statistical Learning Theory (1998)
- F. Cucker & S. Smale: On the mathematical foundations of learning (2001)
- O. Bousquet, S. Bucheron, G. Lugosi: Introduction to statistical learning theory