CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

L19: Learning Theory

Goals of this lecture

- Getting to know learning theory basics

Reading material:

- Shai Shalev-Shwartz & Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms, Chapter 4

Learning Theory

Learning on a training set is fine, but what can we say about generalization?

Possible analysis:

- Vapnik-Chervonenkis theory (often distribution independent and therefore worst case, mostly applicable to binary classification)
- Rademacher complexity
- PAC-Bayes (probably approximately correct)
  - Define a prior P over the function class $\mathcal{F}$
  - Algorithm outputs a posterior Q over the function class $\mathcal{F}$

Learning Algorithm

$$A : \mathcal{D} \rightarrow \mathcal{F}$$

$\mathcal{D} = \{(x^{(i)}, y^{(i)}\}$: Dataset
$\mathcal{F}$: Space of all classifiers

Learning Theory Assumptions:

- There exists a data-generating distribution P_d over $\mathcal{D}$
- Algorithm doesn't know the distribution but sees only $\mathcal{D}$
- The samples in $\mathcal{D}$ are i.i.d.

Learning a predictor:

- Algorithm driven by some learning principle: max-margin, negative log-likelihood, etc.
- Informed by prior knowledge resulting in **inductive bias**

Certifying performance:

- What happens beyond the training set
- Generalization bounds

If a classifier does well on the given dataset, will it do well on other data drawn from P_d?

Losses:

Which losses have we seen so far:

- $\ell(f(x), y) = \delta(f(x) \neq y)$: 0-1 loss
- $\ell(f(x), y) = (y - f(x))^2$: square loss
- $\ell(f(x), y) = \max\{0, (1 - yf(x))\}$: hinge loss
- $\ell(f(x), y) = \log(1 + \exp(yf(x)))$: log loss

Let's define the **bounded** random variable

$$X_i = \delta(f(x^{(i)}) \neq y^{(i)}) = \begin{cases} 1 & \text{if } f(x^{(i)}) \neq y^{(i)} \\ 0 & \text{otherwise} \end{cases}$$

Let's define

$$\begin{aligned} &\text{(empirical error)} \qquad & L_{\mathcal{D}}(f) = & \frac{1}{|\mathcal{D}|} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} X_i \\ &\text{(generalization error)} \qquad & L(f) = & \mathbb{E}_{(x,y) \in P_d}[X] \end{aligned}$$

Generalization gap:

$$L(f) - L_{\mathcal{D}}(f)$$

Goal:

$$L(f) \leq L_{\mathcal{D}}(f) + \epsilon(\cdot)$$

How to get to a bound

$$L(f) \leq L_{\mathcal{D}}(f) + \epsilon(\cdot)$$

Hoeffding's inequality:

- $X_i \in [0, 1]$ i.i.d.
- $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$
- We know

$$P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2e^{-2nt^2}$$

In our case with **union bound**:

$$P(\exists f \in \mathcal{F} : |L(f) - L_{\mathcal{D}}(f)| > t) \leq \sum_{f \in \mathcal{F}} P(|L(f) - L_{\mathcal{D}}(f)| > t) \leq 2|\mathcal{F}|e^{-2|\mathcal{D}|t^2}$$

Our bound:

$$P(\exists f \in \mathcal{F} : |L(f) - L_{\mathcal{D}}(f)| > t) \leq \sum_{f \in \mathcal{F}} P(|L(f) - L_{\mathcal{D}}(f)| > t) \leq 2|\mathcal{F}|e^{-2|\mathcal{D}|t^2}$$

We want this bound to be less than δ, i.e., $2|\mathcal{F}|e^{-2|\mathcal{D}|t^2} \leq \delta$

$$|\mathcal{D}| \geq \frac{\ln(2|\mathcal{F}|) + \ln(\frac{1}{\delta})}{2t^2}$$

Consequently: With probability at least $1 - \delta$ (flip inequality in 'our bound')

$$L(f) \leq L_{\mathcal{D}}(f) + \sqrt{\frac{\ln(2|\mathcal{F}|) + \ln(\frac{1}{\delta})}{2|\mathcal{D}|}}$$

Note: This only works for finite function classes

Observations from

$$L(f) \leq L_{\mathcal{D}}(f) + \sqrt{\frac{\ln(2|\mathcal{F}|) + \ln(\frac{1}{\delta})}{2|\mathcal{D}|}}$$

- Increasing $|\mathcal{D}|$ decreases the second term
- Low empirical error guarantees lower generalization error
- A simple hypothesis space (small $\ln(|\mathcal{F}|)$) decreases generalization error

Note: Many generalizations exist and are covered in other classes

Quiz:

- Goal of learning theory?
- Hoeffding bound?
- Union bound?

Important topics of this lecture

- Understanding how we can bound the generalization gap

Up next:

- Generative modeling