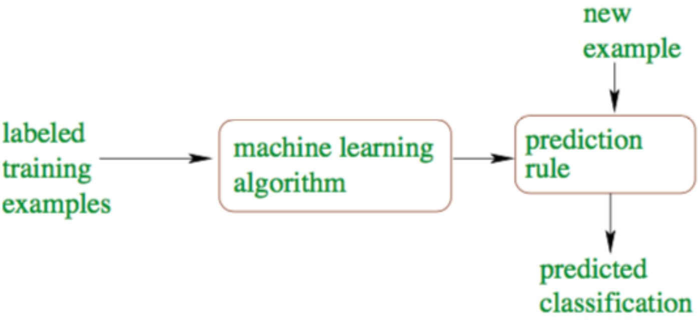
Mathematical techniques in data science

Vu Dinh

Lecture 3: Generalization bounds

February 18th, 2019

Supervised learning problem



Supervised learning: learning a function that maps an input to an output based on example input-output pairs

Supervised learning: standard setting

- 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 $P_{X,Y}$
- Goal: predict the label of a new instance x

Supervised learning: standard setting

- 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

- The function h is an element of some space of possible functions $\mathcal{H}$, usually called the *hypothesis space*
- In order 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

Risk and empirical risk

- With a pre-defined loss function, the "optimal hypothesis" is the minimizer over $\mathcal{H}$ of the risk function

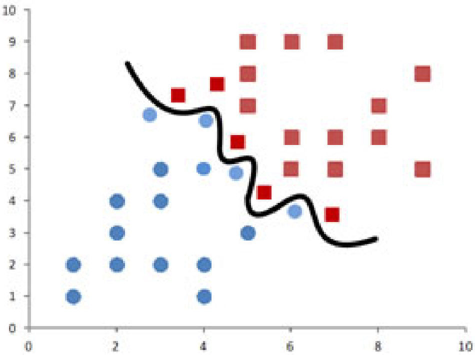
$$R(h) = E_{(X,Y)\sim P}[L(Y, h(X))]$$

- Since P is unknown, the simplest approach is to approximate the risk function by the empirical risk

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

- The empirical risk minimizer (ERM): minimizer of the empirical risk function

Overfitting



Definition

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 $(1-\delta)$.

Exponential moment of bounded random variables

Theorem

For any random variable X, $\epsilon > 0$ and $t > 0$

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

Theorem

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

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

Hoeffding's inequality

Theorem (Hoeffding's inequality)

Let $X_1, X_2, \ldots, X_n$ be i.i.d copy of a random variable $X \in [a, b]$, and $\epsilon > 0$,

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

Generalization bound for finite hypothesis space and bounded loss

Assumption

- the loss function L is bounded, that is

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

- the hypothesis space is a finite set, that is

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

PAC estimate for ERM

Theorem

For any $\delta > 0$ *and* $\epsilon > 0$*, if*

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

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