

Mathematical techniques in data science

Lecture 15: Hoeffding's inequality

What's next?

- Mathematical techniques in data sciences
 - A short introduction to statistical learning theory
 - Linear regression – regularization and feature selection
 - SVM – the kernel trick
 - Random forests — boosting and bootstrapping
- Algorithms and learning contexts
 - PCA and Manifold learning
 - Clustering
 - Selected topics

Supervised learning: standard setting

- Given: a sequence of label data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ sampled (independently and identically) from an unknown distribution $P_{X,Y}$
- Goal: predict the label of new samples (as accurately as possible)
- Question:
 - How to make predictions?
 - What do you mean by “as accurately as possible?”

Hypothesis space

- Given: a sequence of label data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ sampled (independently and identically) from an unknown distribution $P_{\mathcal{X}, \mathcal{Y}}$
- Goal: a learning algorithm seeks a function $h : \mathcal{X} \rightarrow \mathcal{Y}$, 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*
- Usually, this hypothesis space can be indexed by some parameters (often specified by a model or a learning algorithm)

Hypothesis space: logistic regression

- Two classes: 0 and 1
- $x \in \mathbb{R}^d$
- Probability model

$$p_{w,b}(x) = \frac{1}{1 + e^{-w^T x - b}}$$

- Prediction rule $h_{w,b}(x)$
 - If $p_{w,b}(x) > 0.5$, predict $h_{w,b}(x) = 1$
 - If $p_{w,b}(x) \leq 0.5$, predict $h_{w,b}(x) = 0$
- Hypothesis space

$$\mathcal{H} = \{h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

Loss function

- In order to measure how well a function fits the data, a *loss function*

$$L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^{\geq 0}$$

is defined

- For regression:

$$L(h(x), y) = [h(x) - y]^2$$

- For classification:

$$L(h(x), y) = \begin{cases} 0, & \text{if } h(x) = y \\ 1 & \text{otherwise} \end{cases}$$

Loss function

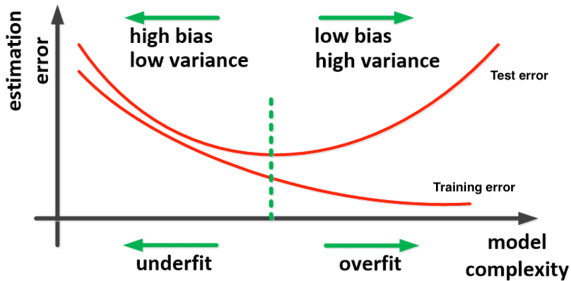
- The function h is an element of some space of possible functions \mathcal{H} , usually called the *hypothesis space*
- In order to measure how well a function fits the data, a *loss function*

$$L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^{\geq 0}$$

is defined

- It is straightforward that we want to have a hypothesis with minimal loss
- Question: minimal loss on what?

Underfitting/Overfitting



Risk function

- Assumption: The future samples will be obtained from the same distribution $P_{X,Y}$ of the training data
- With a pre-defined loss function, the risk function is defined as

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

- The “optimal hypothesis”, denoted by h^* in this lecture, is the minimizer over \mathcal{H} of the risk function

$$h^* = \arg \min_{h \in \mathcal{H}} R(h)$$

Empirical risk

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

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

- Rationale: The law of large number – If the random variables Z_1, Z_2, \dots, Z_n are drawn independently from the same distribution P_Z , then

$$\frac{Z_1 + Z_2 + \dots + Z_n}{n} \approx E[Z]$$

ERM

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

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

- Rationale: $R_n(h) \approx R(h)$
- In this lecture, we use the notation \hat{h}_n to denote the ERM
- We hope that

$$R(\hat{h}_n) \approx R(h^*)$$

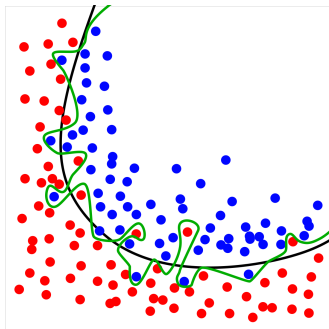
- Note: \hat{h}_n is random, while h^* is a fixed hypothesis

Failure of ERM

We hope that

$$R(\hat{h}_n) \approx R(h^*),$$

but in general, this might not be true if the hypothesis space \mathcal{H} is too large



Failure of ERM

- We hope that

$$R(\hat{h}_n) \approx R(h^*),$$

but in general, this might not be true if the hypothesis space \mathcal{H} is too large

- Question: How large is too large?
- We need to be able to quantify/control the difference between $R(\hat{h}_n)$ and $R(h^*)$

Modes of estimations

- Analysis

$$\lim_{n \rightarrow \infty} x_n = x$$

- Numerical analysis

$$\|x_n - x\| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \quad \text{or} \quad \|x_n - x\| \leq \frac{C}{\sqrt{n}}$$

- PAC (Probably Approximately Correct) learning

$$\|x_n - x\| \leq C(\delta) \frac{1}{\sqrt{n}}$$

with probability at least $1 - \delta$

PAC learning

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

$$P \left[R(\hat{h}_n) - R(h^*) > \epsilon \right] < \delta.$$

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

Probability inequalities

Markov inequality

Theorem (Markov inequality)

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

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

Markov inequality

Theorem

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

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

Exponential moment of bounded random variables

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

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

Corollary:

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