Mathematical techniques in data science

Vu Dinh

Departments of Mathematical Sciences
University of Delaware

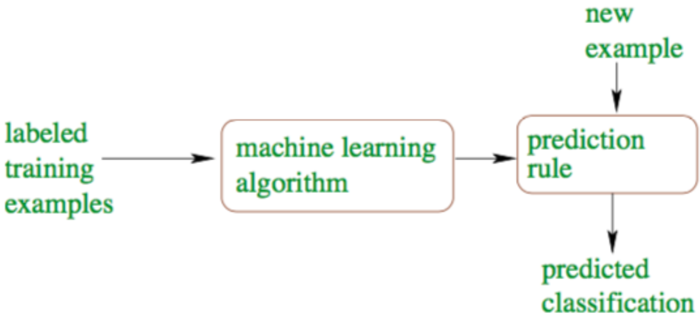
February 15th, 2019

Tentative schedule

Week	Chapter
1	Chapter 2: Intro to statistical learning
2	Chapter 4: Classification
3	Chapter 9: Support vector machine and kernels
4, 5	Chapter 3: Linear regression
6	Chapter 8: Tree-based methods + Random forrest
7	Neural network
8	
9	Bootstrap and cross-validation + Bayesian methods + UQ
10	Clustering: K-means –> Spectral Clustering
11	PCA –> Manifold learning
12, 13	Reinforcement learning/Online learning/Active learning
14	Project presentation

An introduction to statistical learning

Diagram of a typical 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

Example

- MNIST dataset

- You are provided a dataset containing images (16 x 16 grayscale images) of digits.
- Each image contains a single digit.
- Each image is labelled with the corresponding digit
- Can think of each image as a vector in $X \in \mathbb{R}^{256}$ and the label as a scalar $Y \in \{0, 1, \ldots, 9\}$
- Goal: learn to identify/predict digits

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*.

- 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

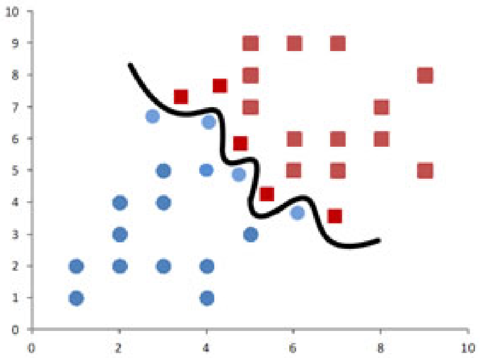
Learning complexity

- The central idea of machine learning is that *the past informs the future*, which means that in general

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

  uniformly on $\mathcal{H}$
- This requires that the hypothesis space $\mathcal{H}$ needs not be too large, otherwise overfitting will occur

Overfitting



Choosing a proper hypothesis space plays a central role in statistical learning

PAC learning

PAC learning

- Analysis

$$\lim_{k\to\infty} x_k = x$$

- Numerical analysis

$$\|x_n - x\| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

- 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

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

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

Markov inequality

Theorem (Markov inequality)

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

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

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

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