Mathematical techniques in data science

Lecture 14: A short introduction to statistical learning theory
– Hypothesis spaces and loss functions–

Reminders

- Office hours:
  - MW 3pm-4pm, Ewing Hall 312
  - By appointments
- Homework 2: due 03/21 EOD

Where are we?

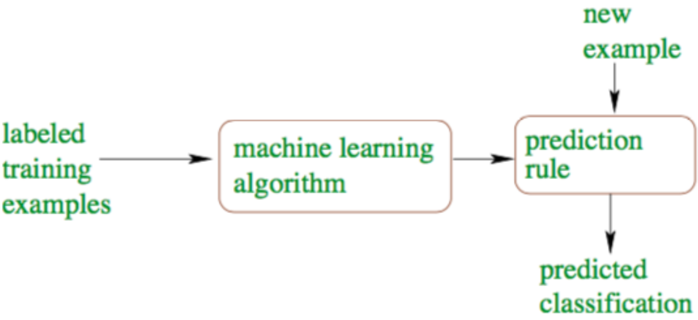
- Algorithms
  - Intros to classification
  - Overfitting and underfiting
  - Nearest neighbors
  - Logistic regression
  - Feed-forward neural networks
- Codings
  - Numpy, matplotlib, sklearn
  - Reading sklearn documentations
  - Pre-process inputs (i.e., numpy.shape())
  - Data simulations (by hand or using built-in functions in sklearn)
  - Data splitting
  - Train models; making prediction; evaluate models

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

A short introduction to statistical learning theory

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 new samples (as accurately as possible)

Example

- MNIST dataset

- 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 (as accurately as possible)

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 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ sampled (independently and identically) from an unknown distribution $P_{X,Y}$
- 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

- 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*

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

  is defined

Loss function: examples

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

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

  is defined

- For regression:

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

- For classification:

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

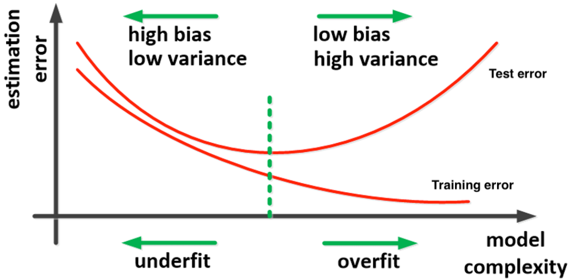
Loss function

- 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*

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

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

Underfiting/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

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