Statistical Learning

From Kroese

Some important aspects of data science:

- structuring data
- visualizing data
- mathematical analysis of data (the main challenge)

When the goal is to interpret the model and quantify the uncertainty in the data, this analysis is usually referred to as *statistical learning*.

When the emphasis in on making predictions using large-scale data, the it is common to speak about machine learning or data mining.

Two major goals for modeling data:

- to accurately predict some future quantity of interest
- to discover unusual or interesting patterns in the data

To achieve these goals, one must rely on knowledge from three important pillars of the mathematical sciences:

- Function approximation
- Optimization
- Probability and statistics

1 *Supervised and unsupervised learning*

Given an input or feature vector $\mathbf{x}$, predict an output or *response* variable y.

The prediction is encoded in a mathematical function g, called the prediction function, which takes an input $\mathbf{x}$ and outputs a guess $g(\mathbf{x})$ for y, denoted by $\widehat{y}$.

Example (Polynomial regression)

Input: u

Prediction function: g

Function space: $\mathcal{H}$

Data points: (u_i, y_i), $i = 1, \ldots, n$ drawn from iid random points (U_i, Y_i), where U_i are uniformly distributed on the invterval $(0, 1)$ and given $U_i = u_i$, the random variable Y_i has a normal distribution with expectation $10 - 140u_i + 400u_i^2 - 250u_i^3$ and variance $l^* = 25$

Using a squared-error loss, the optimal prediction function $h^*(u) = \mathbb{E}[Y|U = u]$ is:

$$h^*(u) = 10 - 140u + 400u^2 - 250u^3$$

To obtain a good estimate of $h^*(u)$ based on the training set $\tau = \{(u_i, y_i), i = 1, \ldots, n\}$ we minimize the outcome of the training loss:

$$l_\tau(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h(u_i))^2 \tag{1}$$

over a suitable set $\mathcal{H}$ of the candidate functions.
Let's take the set $\mathcal{H}_p$ of polynomial functions in u or order $p - 1$:

$$h(u) := \beta_1 + \beta_2 u + \beta_3 u^3 + \cdots + \beta_p u^{p-1} \tag{2}$$

for $p = 1, 2, \ldots$ and parameter vector:

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^\top$$