Machine Learning

1 Terminology

- Data generating process: (X, Y) is a $(p+1)-$dimensional random vector with joint distribution $\mathrm{P}(x, y)$.
  - Input vector: $X \in D \subset \mathbb{R}^p$.
  - Output vector: $Y \in E \subset \mathbb{R}$.
  - Data: Given the sample $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$ following the distribution $\mathrm{P}(x, y)$, The training data or text data $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ consist of the realization values of the sample.
- Objective: Find a decision function $\hat{f}$ to minimize the expected loss

$$\min_{\hat{f} \in \mathcal{F}} \mathrm{E}(L(Y, \hat{f}(X)))$$

  - Decision function: $\hat{f} : \mathbb{R}^p \supset D \longrightarrow \mathbb{R}$ serves to produce the prediction $\hat{y} = \hat{f}(x)$ of Y, provided a specified value x of X.
  - Loss function: $L(Y, \hat{f}(X))$ normally has the form of

$$L_2 = (Y - \hat{f}(X))^2 \text{ or } L_1 = |Y - \hat{f}(X)|.$$

  - Hypothesis space: $\mathcal{F}$ is a collection of all decision functions $\hat{f}$ to be selected. In some cases, we suppose that as a candidate $\hat{f}$ can be specified by several parameters. Thus $\mathcal{F} = \{\hat{f}_\theta : Y = \hat{f}_\theta(X),\ \theta \in \mathbb{R}^n\}$ can be described by the parametric space $\Theta = \{\theta : Y = \hat{f}_\theta(X),\ \theta \in \mathbb{R}^n\}$.
- Optimization strategies:
  empirical risk minimization:

$$\min_{\hat{f} \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

  structural risk minimization:

$$\min_{\hat{f} \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$