# Machine Learning

# Chapter 1

# Introduction

## 1.1 Terminology and framework

- 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: $Y \in G \subset \mathbb{R}$.
  - 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 optimal decision function $\hat{f}$ to minimize the expected prediction loss (EPE)

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

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

$$L_2 = (Y - f(X))^2 \quad \text{or} \quad L_1 = |Y - f(X)| \quad \text{or} \quad L_I = 1_{Y \neq f(X)}.$$

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

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

  structural risk minimization:

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

## 1.2 Squared error loss

### 1.2.1 Quantitative output variables

Let $Y \in \mathbb{R}$ be quantitative variable. And we take the most common and convenient loss function, squared error loss

$$L(Y, f(X)) = (Y - f(X))^2.$$

That leads to the problem of minimizing the expected prediction error

$$\min_{f \in \mathcal{F}} \mathrm{E}[(Y - f(X))^2] = \min_{f \in \mathcal{F}} \int (y - f(x))^2 \, \mathrm{dP}(x, y).$$

Note that

$$\mathrm{E}[(Y - f(X))^2] = \mathrm{E}[\mathrm{E}[(Y - f(X))^2 | X]] = \int \mathrm{E}[(Y - f(x))^2 | X = x] \, \mathrm{dP}(x).$$

It suffices to minimize EPE pointwise, that is,

$$\begin{aligned}
&\min_{f(x) \in \mathbb{R}} \mathrm{E}[(Y - f(x))^2 | X = x] \\
&= \min_{c \in \mathbb{R}} \mathrm{E}[(Y - c)^2 | X = x] \\
&= \min_{c \in \mathbb{R}} \mathrm{Var}[Y - c | X = x] - (\mathrm{E}[(Y - c) | X = x])^2 \\
&= \min_{c \in \mathbb{R}} \mathrm{Var}[Y | X = x] - (\mathrm{E}[Y | X = x] - c)^2.
\end{aligned}$$

We see the optimal solution is

$$\hat{f}(x) = \mathrm{E}[Y | X = x],$$

Thus the best prediction of $Y$ at any point $X = x$ is the conditional expectation, when best is measured by average squared error.
Next we are developing effective methods to estimate the conditional expectation $\mathrm{E}[Y | X = x]$.

### 1.2.2 Categorical output variable

Assume that $Y \in G$ is a categorical variable and that the set of possible classes $G = \{G_1, G_2, \cdots, G_K\}$. This time the 0–1 loss function

$$L(Y, \hat{f}(X)) = 1_{Y \neq \hat{f}(X)} = \begin{cases} 0, & Y = \hat{f}(X), \\ 1, & Y \neq \hat{f}(X), \end{cases}$$

is adopted for simplification. Likewise it suffices to minimize EPE pointwise.

$$\begin{aligned}
&\min_{\hat{f}(x) \in G} \mathrm{E}[1_{Y \neq \hat{f}(x)} | X = x] \\
&= \min_{g \in G} \mathrm{E}[1 - 1_{Y = g} | X = x] \\
&= \min_{g \in G} 1 - \mathrm{P}(Y = g | X = x)
\end{aligned}$$

And the optimal solution is

$$\hat{f}(x) = \max_{g \in G} \mathrm{P}(Y = g | X = x)$$

## 1.3 Nearest-neighbor methods

Nearest-neighbor methods use those observations in the training set $T$ closest in input space to $x$ to estimate the aforementioned conditional expectation $\mathrm{E}[Y|X = x]$. Specifically, the $k$-nearest neighbor fit for $\hat{Y}$ is defined as follows:

$$\hat{Y} = \hat{f}(x) = \frac{1}{k} \sum_{i:x_i \in N_k(x)} y_i,$$

where $N_k(x)$ is the neighborhood of $x$ defined by the $k$ closest points $x_i$ in the training sample. Closeness implies a metric, which for the moment we assume is Euclidean distance.