I. Supervised Learning: introduction

The general goal of supervised learning is to learn decision rules from labeled examples. The examples are denoted by

\[ X_1, \ldots, X_n \in \mathcal{X} \] (feature space)

while the labels are

\[ Y_1, \ldots, Y_n \in \mathcal{Y} \] (label set)

It is assumed that \((X_i, Y_i)\) are independent random variables drawn from a distribution \(P\). This distribution is unknown. The aim is to design a prediction rule,

\[ g: \mathcal{X} \rightarrow \mathcal{Y} \]

such that for every "new" pair \((X, Y)\) drawn from \(P\), \(g(X)\) is very likely to be a good prediction of \(Y\).

Example 1. (Character recognition)

Each example \(X_i\) corresponds to a digital image of a digit \(0, 1, 2, \ldots, 9\) (the interested reader may have a look on the MNIST dataset). Pay attention \(X_i\) is an image representing a digit, not a digit by itself.
Usually $X_i \in \{0,1,\ldots,9\}$ and $Y_i \in \{0,1,\ldots,9\}$.

The goal is to find an automatic rule that takes as input an image and provides as output an element of $Y = \{0,1,\ldots,9\}$.

**Example 2** (Prediction of stock option prices).

Let $P_t$ be the price of a stock option at time $t$.

Our goal is to use the historical data $(P_{t-k+1},\ldots,P_t)$ in order to predict the highest value in the near future: $\max_{1 \leq j \leq 30} P_{t+j}$ (highest value of the next 30 days).

So here $Y = \max_{1 \leq j \leq 30} P_{t+j} \in \mathbb{R}_+$

$X = (P_{t-k+1},\ldots,P_t) \in \mathbb{R}^k$

Usually in this problem, it is better to transform these variables as follows:

$Y = \max_{1 \leq j \leq 30} (P_{t+j} - P_t)/P_t \in \mathbb{R}$

$X = \left( \frac{P_{t-k+1}}{P_{t-1}}, \frac{P_{t-k+1}}{P_{t-k}}, \ldots, \frac{P_{t-k+1}}{P_{t-k}} \right) \in \mathbb{R}^k$

Considering different stock options and different time periods, we get our training sample $(X_1,Y_1),\ldots,(X_n,Y_n)$

This sample can be used to infer a prediction rule.

Ⅱ Bayes Predictor

The setting: $P$ is a probability on $\mathbb{X} \times \mathbb{Y}$

$(X_i,Y_i) \overset{iid}{\sim} P \quad i = 1,\ldots,n$
We look for a prediction function 
\[ g : \mathcal{X} \to \mathcal{Y}. \]
To quantify the quality of \( g \), we introduce a loss function
\[ l : \mathcal{Y} \times \mathcal{Y}^* \to \mathbb{R}_+ \]
Here \( l(y, y') \) corresponds to the loss incurred when \( y \) is predicted by \( y' \). Generally, the loss function satisfies the relation \( l(y, y) = 0 \) \( \forall y \in \mathcal{Y} \).

**Example 1** (Binary classification)
Here, \( \mathcal{X} \) is arbitrary and \( \mathcal{Y} = \{0, 1\} \) or \( \mathcal{Y} = \{-1, +1\} \).
The usual loss in this setting is the 0-1 loss
\[ l(y, y') = I(y \neq y') \]
The risk of a prediction function \( g \) is then
\[ R_P(g) = \mathbb{E}[l(Y, g(X))] = \mathbb{P}(Y \neq g(X)) \]

**Example 2** (Least-squares regression)
The set \( \mathcal{X} \) is still arbitrary and \( \mathcal{Y} = \mathbb{R} \).
The squared loss is \( l(y, y') = (y - y')^2 \) and the risk is
\[ R_P(g) = \mathbb{E}[(Y - g(X))^2] \]

**DEF.** We call the Bayes rule any prediction function 
\[ g^* : \mathcal{X} \to \mathcal{Y} \]
satisfying
\[ g^* \in \text{arg min}_g R_P(g) \quad (\Rightarrow R_P(g^*) \leq R_P(g) \quad \forall g) \]
At a heuristic level, the Bayes rule is the best prediction function that we would use if we were given the probability \( P \). Since \( P \) is unknown, we cannot use \( g^* \) directly.
**Theorem**

Let $P$ be a probability on $X \times Y$ and $R_{P}(g) = \mathbb{E}[\ell(Y, g(x))]$.

a) The Bayes rule $g^*_P$ can be computed by

$$g^*_P(x) \in \arg\min_{a \in Y} \mathbb{E} [\ell(Y, a) \mid X = x] \quad \forall x \in X$$

b) In the problem of regression with least-squares loss

$$g^*_P(x) = \mathbb{E}[Y \mid X = x] \quad \forall x \in X$$

c) In the problem of binary classification with $Y = \{0, 1\}$,

$$g^*_P(x) = \mathbb{1} \left( \eta(x) > \frac{1}{2} \right) \quad \forall x \in X$$

where $\eta(x) = \mathbb{E}[Y \mid X = x] = P(Y = 1 \mid X = x)$.

**Proof.** According to the total probabilities formula

$$P(dx, dy) = P(dy \mid X = x) \cdot P_x(dx)$$

where $P_x(dx)$ is the marginal distribution of $X$.

a) Therefore,

$$R_{P}(g) = \mathbb{E}[\ell(Y, g(x))] = \int_{X \times Y} \ell(y, g(x)) P(dx, dy)$$

$$= \int_{X} \left( \int_{Y} \ell(y, g(x)) P(dy \mid X = x) \right) P_x(dx)$$

$$= \int_{X} \mathbb{E}[\ell(Y, g(x)) \mid X = x] P_x(dx)$$

$$\geq \int_{X} \min_{a} \mathbb{E}[\ell(Y, a) \mid X = x] P_x(dx)$$

$$= \int_{X} \mathbb{E}[\ell(Y, g^*(x)) \mid X = x] P_x(dx)$$

$$= R_{P}(g^*)$$

This implies that $R_{P}(g) \geq R_{P}(g^*)$ for every $g$,

which means that $g^*$ is the Bayes rule.
b) When $l(y, g(x)) = (y - g(x))^2$, applying a) we get

$$g^*(x) = \arg \min_{a \in \mathbb{R}} E \left[ \frac{(Y-a)^2}{\mathbb{P}(Y \mid X=x)} \right]$$

We have $F(a) = E[Y^2 \mid X=x] - 2aE[Y \mid X=x] + a^2$.

The minimum of this function is attained when

$$a = E[Y \mid X=x].$$

c) For $l(y, a) = 1(y \neq a)$ we have

$$\arg \min_{a \in \{0, 1\}} E\left[1(Y=a) \mid X=x\right]$$

$$= \arg \min_{a \in \{0, 1\}} P(Y=a \mid X=x)$$

$$= \arg \max_{a \in \{0, 1\}} P(Y=a \mid X=x)$$

$$= \begin{cases} 1, & \text{if} \ P(Y=1 \mid X=x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

III) Empirical risk minimization

$(X_i, Y_i) \sim P \quad \forall : \mathcal{X} \rightarrow \mathcal{Y} \quad R_p(g) = E[l(Y, g(X))]$

We want now to find $g$ such that $R_p(g)$ is small without using the probability $P$.

The main idea is that when $n$ is large the empirical risk

$$\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, g(X_i))$$

is a good approximation of $R_p(g)$. Indeed, according to the central limit theorem.
\[ \hat{R}_n(g) - R_P(g) \approx \frac{\mathbb{E}(g)}{\sqrt{n}} \]

where \( \mathbb{E}(g) \sim N(0, \sigma^2) \). However, this relation is true only for a fixed \( g \). If \( G \) is a very wide class of functions, the quantity

\[ \sup_{g \in G} (\hat{R}_n(g) - R_P(g)) \]

does not necessarily go to 0 when \( n \to +\infty \).

**DEF.** Given a set of candidate prediction functions, \( G \), we call empirical risk minimizer (ERM) the function

\[ \hat{g}_n \in \arg \min_{g \in G} \hat{R}_n(g). \]

The choice of the set \( G \) is of central importance. This is clear from the following decomposition:

\[ R_P(\hat{g}_n) - R_P(g^*) = R_P(\hat{g}_n) - R_P(\hat{g}_n^*) + R_P(\hat{g}_n^*) - R_P(g^*) \]

where \( \hat{g}_n^* \in \arg \min_{g \in G} R_P(g) \).

It is clear that both \( T_1 \) and \( T_2 \) are \( > 0 \). In addition \( T_1 \) increases when \( G \) becomes larger, whereas \( T_2 \) decreases when \( G \) increases.

- \( T_1 \) is called statistical error.
- \( T_2 \) is called bias or approximation error.

When \( T_1 \) is too small and \( T_2 \) is too large, we say that \( \hat{g}_n \) underfits. When \( T_2 \) is too small and \( T_1 \) too large, then \( \hat{g}_n \) overfits.