1 Introduction

This course aims to give an introduction to the theory of learning from data in a probabilistic framework. In the spirit of Vapnik that *nothing is more practical than a good theory*, we will examine the theoretical aspects of statistical learning. Although learning from data includes a variety of topics such as regression, classification, density estimation, and clustering in general, we will mainly focus on the classification or pattern recognition. Concepts, insights, and theoretical results discussed for this specific type of learning will be useful and broadly applicable to understanding other sets of problems.

First, the binary classification problem is generally formulated.

- **Definition: Training data**, $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$ where the predictors, $X_i \in \mathcal{X} = \mathbb{R}^d$ and the class labels, $Y_i \in \mathcal{Y} = \{0, 1\}$.
  Here (X_i, Y_i)'s are i.i.d with some unknown distribution $\mathcal{P}_{X,Y}$ ($\mathcal{P}$ for brevity). We will use $P(\cdot)$ as a generic notation for any *probability* function, where the underlying probability space or the concerned random variable will be either obvious from the context or will be written explicitly as $P_{Y|X}(\cdot)$ or $P(Y|X = x)$.

- **Definition: Decision rule** f. The goal, in a broader sense, is to find a map $f : \mathcal{X} \to \mathcal{Y}$ based on $\mathcal{D}_n$, which can be generalized to future cases $(X, Y) \sim \mathcal{P}_{X,Y}$. Consequently, such a map f depends on the training data $\mathcal{D}_n$ and $f(x)$ is a random variable. Though the term "decision rule" (or "decision function") can be associated to any decision theoretic setup, for our purpose, we will use it interchangeably with the term "classification rule" (or "classification function").

- **Definition: Loss function**, $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ in general. For a choice of f, we need a performance criterion, and the loss function L defines such a criterion. The most commonly used loss for classification problems is the 0-1 loss (or the *misclassification error*) given by
$$L(f(x), y) := I(f(x) \neq y).$$

- **Definition: Risk**, the expected loss,
$$R(f) := E_{\mathcal{P}} L(f(X), Y) = \mathcal{P}(f(X) \neq Y).$$
It is the theoretical probability of error incurred by the classification rule f. Ideally we want to find f with its risk $R(f)$ as small as possible. The **Bayes risk** is defined as the smallest risk achievable by any measurable decision function, that is,
$$R^* := \inf_{\text{all measurable } f} R(f),$$
where the infimum is taken over all measurable functions $f : \mathcal{X} \to \mathcal{Y}$.

- **Definition**: **Conditional class probability**, $\eta : \mathcal{X} \to [0, 1]$. Given x, the probability that the datum belongs to class one: $\eta(x) := P(Y = 1|X = x)$.

- **Definition**: The Bayes decision function or **Bayes decision rule**,
$$f^*(x) := I\left(\eta(x) \geq \frac{1}{2}\right) = \begin{cases} 1 & \text{if } \eta(x) \geq 1/2, \\ 0 & \text{otherwise.} \end{cases} . \tag{1.1}$$

Theorem 1. *For any **binary** classification rule $f : \mathcal{X} \to \mathcal{Y}$, with 0-1 loss,*

$$R(f) - R(f^*) = E_X\{I(f(X) \neq f^*(X)) \cdot |2\,\eta(X) - 1|\}$$

Remark 1. Before going through the proof, let us see what $R(f) - R(f^*) \geq 0$ means: For any measurable f, the increase in risk relative to f^* depends on how far $\eta(x)$ is from $1/2$ on the set of x's where f and f^* disagree. In other words, if $\eta(x) = 1/2, \forall\ x \in \{x \in \mathcal{X} : f(x) \neq f^*(x)\}$, then there is no excess risk. That is, the rule f is equivalent to the best possible (Bayes) rule f^*, in terms of the average consequence (risk).

Proof. Note that $Y, f(\cdot), f^*(\cdot) \in \{0, 1\}$ and

$$E_{Y|X=x} I(Y = f^*(x)) = P_{Y|X}(Y = f^*(x)) = \begin{cases} \eta(x) & \text{if } \eta(x) \geq 1/2, \text{ [using (1.1)]} \\ 1 - \eta(x) & \text{otherwise} \end{cases}$$
$$= 1/2 + |\eta(x) - 1/2|.$$

Also, remember the conditional expectation formula:

$$E_{X,Y}\, g(X, Y) = E_X E_{Y|X}\, g(X, Y).$$

$$\begin{aligned}
\text{So, } R(f) - R(f^*) &= E_{\mathcal{P}}\Big\{ I(f(X) \neq Y) - I(f^*(X) \neq Y) \Big\} \\
&= E_{\mathcal{P}}\Big\{ I(f(X) \neq f^*(X)) \Big(I(f(X) \neq Y) - I(f^*(X) \neq Y) \Big) \Big\} \\
&= E_{\mathcal{P}}\Big\{ I(f(X) \neq f^*(X)) \Big(I(f^*(X) = Y) - I(f^*(X) \neq Y) \Big) \Big\} \\
&= E_{\mathcal{P}}\Big\{ I(f(X) \neq f^*(X)) \Big(2\, I(f^*(X) = Y) - 1 \Big) \Big\} \\
&= E_X\Big\{ I(f(X) \neq f^*(X)) \Big(2\, E_{Y|X} I(f^*(X) = Y) - 1 \Big) \Big\} \\
&= E_X\Big\{ I(f(X) \neq f^*(X)) \Big| 2\,\eta(X) - 1 \Big| \Big\}.
\end{aligned}$$

□

Though it may seem obvious, but it is worthwile to note the following result:

Result 1. $R^* = R(f^*) = E_X\{\min\{\eta(X), 1 - \eta(X)\}\}$.

Proof. $R(f) - R(f^*) \geq 0$ for all measurable $f \Rightarrow \boxed{R^* = R(f^*)}$ i.e. f^* is indeed the Bayes (best possible) rule achieving the lowest possible risk. Now for the other equality,

$$P(Y \neq f^*(X) | X = x) = 1 - P(Y = f^*(X) | X = x) = \begin{cases} 1 - \eta(x) & \text{if } \eta(x) \geq 1/2 \\ \eta(x) & \text{if } \eta(x) < 1/2. \end{cases}$$
$$= \min\{\eta(x), 1 - \eta(x)\}$$

$$\Rightarrow \boxed{R^* = \mathcal{P}(f^*(X) \neq Y) = E_X\{\min\{\eta(X), 1 - \eta(X)\}\}.} \tag{1.2}$$

□

Thus, if in some idealistic situation, $\eta(x) \in \{0,1\}\ \forall\ x \in \mathcal{X}$, then a perfect separation is possible with $R^* = 0$! Note that the Bayes decision rule requires the complete knowledge of the underlying probability distribution or $\eta(x)$. In general, the Bayes error rate indicates how difficult the given classification problem is.

In the lack of the complete knowledge of η, which is the case in practice, an important question is how to build a classification rule from the data with a minimum error rate. There are two different approaches. One can construct a probability model for the data first and use it for classification. Or one may aim to minimize the error rate directly.

Probability Modelling: Mimic the Bayes rule $f^*(x) = I(\eta(x) \geq 1/2)$ by estimating η by $\hat{\eta} : \mathcal{X} \to [0,1]$ with the data (e.g. logistic regression) and using the plug-in decision function

$$\hat{f}(x) = I(\hat{\eta}(x) \geq 1/2).$$

The following theorem gives some measure of quality of this estimate $\hat{f}$ by comparing its risk with that of the Bayes rule f^*.

Theorem 2.

$$R(\hat{f}) - R(f^*) \leq 2\, E|\hat{\eta}(X) - \eta(X)|.$$

Proof. Note that $\hat{f}(x), f^*(x) \in \{0,1\}$. So, if for some $x \in \mathcal{X}$, $\hat{f}(x) \neq f^*(x)$ then $(\hat{\eta}(x) - 1/2)$ and $(\eta(x) - 1/2)$ have different signs. In this case,

$$|\hat{\eta}(x) - \eta(x)| = |\{\hat{\eta}(x) - 1/2\} - \{\eta(x) - 1/2\}| = |\hat{\eta}(x) - 1/2| + |\eta(x) - 1/2| \geq |\eta(x) - 1/2|.$$

Therefore

$$\begin{aligned} R(\hat{f}) - R(f^*) &= E\{I(\hat{f}(X) \neq f^*(X))\, |2\,\eta(X) - 1|\} \quad \text{[using Thm. 1]} \\ &\leq 2\, E\{I(\hat{f}(X) \neq f^*(X))\, |\hat{\eta}(X) - \eta(X)|\} \\ &\leq 2\, E|\hat{\eta}(X) - \eta(X)|. \end{aligned}$$

□

Remark 2. This theorem implies that estimating $\eta(x)$ by $\hat{\eta}(x)$ accurately yields nearly a minimal risk. Yet, it is not necessary to estimate $\eta(x)$ for finding a rule with a small risk.

Direct Risk Minimization: Find a decision rule that minimizes the error rate $R(f)$ directly without estimating $\eta(x)$.

Since $\mathcal{P}_{X,Y}$ is unknown in practice, we would estimate the probability of error $R(f)$ by the **empirical risk**

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} I(f(X_i) \neq Y_i), \tag{1.3}$$

for any classification rule f. The aim here is to find an f that minimizes the empirical error rate $R_n(f)$ over $\mathcal{F}$, a class of candidate classification rules. As an example, consider the collection of linear classifiers,

$$\mathcal{F} = \{f(x) = I(w'x + b \geq 0) : w \in \mathbb{R}^d \text{ and } b \in \mathbb{R}\}.$$

The classification boundary of each classifier in $\mathcal{F}$ is a hyperplane in $\mathbb{R}^d$.

Emphasizing the dependence of the best rule f on the data, let $f_n^*(x) := f_n^*(x; \mathcal{D}_n)$ be a map $\mathcal{X} \to \mathcal{Y}$ based on $\mathcal{D}_n$ with the minimum empirical error rate in the class. Note that $f_n^*(X)$ is a random function and

$$R(f_n^*) = E_{X,Y} I(f_n^*(X) \neq Y) = P_{X,Y}(f_n^*(X) \neq Y | \mathcal{D}_n).$$

is also a random variable. We will delve into a theory of empirical risk minimization in the following section.