MACHINE LEARNING: THEORY AND APPLICATIONS

Second Lecture, April 8, 2016

American University of Armenia

(I) Supervised Learning: introduction

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

$$X_1, \dots, X_n \qquad \in \mathcal{X} \quad (\text{feature space})$$

while the labels are

$$Y_1, \dots, Y_n \qquad \in \mathcal{Y} \quad (\text{label set})$$

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

$$g: \mathcal{X} \to \mathcal{Y}$$

such that for every "new" pair (X, Y) drawn from $\mathbb{P}$, $g(X)$ 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 $0, 1, 2, \dots, 9$ (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 $X_i \in \{0,1\}^p$ and $Y_i \in \{0,1,\dots,9\}$.
The goal is to find an automatic rule that takes as input an image and provides as output an element of $\mathcal{Y} = \{0,1,\dots,9\}$.

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 $(P_{t-k+1},\dots,P_t)$ in order to predict the highest value in the near future: $\max_{1\le j\le 30} P_{t+j}$ (highest value of the next 30 days).

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

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

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

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

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

Considering different stock options and different time periods, we get our training sample

$(X_1,Y_1),\dots,(X_n,Y_n)$

This sample can be used to infer a prediction rule.

II Bayes Predictor

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

$$(X_i,Y_i) \overset{iid}{\sim} \mathbb{P} \qquad i=1,\dots,n$$

We look for a prediction function

$$g : \mathcal{X} \to \mathcal{Y}.$$

To quantify the quality of g, we introduce a loss function

$$\ell : \mathcal{Y} \times \mathcal{Y}^{\circ} \longrightarrow \mathbb{R}_+$$

Here $\ell(y, y')$ corresponds to the loss incurred when y is predicted by y'. Generally, the loss function satisfies the relation $\ell(y, y) = 0 \quad \forall y \in \mathcal{Y}$.

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

$$\ell(y, y') = \mathbb{1}(y \neq y').$$

The risk of a prediction function g is then

$$R_P(g) = \mathbb{E}\left[\ell(Y, g(X))\right] = \mathbb{P}\left(Y \neq g(X)\right)$$

Example 2 (Least-squares regression)

The set $\mathcal{X}$ is still arbitrary and $\mathcal{Y} = \mathbb{R}$.

The squared loss is $\ell(y, y') = (y - y')^2$ and the risk is

$$R_P(g) = \mathbb{E}\left[(Y - g(X))^2\right].$$

DEF. We call the Bayes rule any prediction function $g^* : \mathcal{X} \to \mathcal{Y}$ satisfying

$$g^* \in \arg\min_g R_P(g) \quad \left(\iff R_P(g^*) \leq R_P(g) \ \forall g\right)$$

At a heuristic level, the Bayes rule is the best prediction function that we would use if we were given the probability $\mathbb{P}$. Since $\mathbb{P}$ is unknown, we can not use g^* directly.

THEOREM

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

a) The Bayes rule $g^*_{\mathbb{P}}$ can be computed by

$$g^*_{\mathbb{P}}(x) \in \arg\min_{a\in\mathcal{Y}} \mathbb{E}\left[\ell(Y,a)\mid X=x\right] \quad \forall x\in\mathcal{X}.$$

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

$$g^*_{\mathbb{P}}(x)=\mathbb{E}\left[Y\mid X=x\right] \quad \forall x\in\mathcal{X}$$

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

$$g^*_{\mathbb{P}}(x)=\mathbb{1}\left(\eta(x)>1/2\right) \quad \forall x\in\mathcal{X}$$

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

Proof. According to the total probabilities formula

$$\mathbb{P}(dx,dy)=\mathbb{P}(dy\mid X=x)\cdot\mathbb{P}_X(dx)$$

where $\mathbb{P}_X(dx)$ is the marginal distribution of X.

a) Therefore,

$$\begin{aligned}
R_{\mathbb{P}}(g)=\mathbb{E}\left[\ell(Y,g(X))\right] &= \int_{\mathcal{X}\times\mathcal{Y}} \ell(y,g(x))\,\mathbb{P}(dx,dy)\\
&= \int_{\mathcal{X}}\left(\int_{\mathcal{Y}}\ell(y,g(x))\,\mathbb{P}(dy\mid X=x)\right)\mathbb{P}_X(dx)\\
&= \int_{\mathcal{X}} \mathbb{E}\left[\ell(Y,g(x))\mid X=x\right]\mathbb{P}_X(dx)\\
&\geq \int_{\mathcal{X}} \min_a \mathbb{E}\left[\ell(Y,a)\mid X=x\right]\mathbb{P}_X(dx)\\
&= \int_{\mathcal{X}} \mathbb{E}\left[\ell(Y,g^*(x))\mid X=x\right]\mathbb{P}_X(dx)\\
&= R_{\mathbb{P}}(g^*)
\end{aligned}$$

This implies that $R_{\mathbb{P}}(g)\geq R_{\mathbb{P}}(g^*)$ for every g, which means that g^* is the Bayes rule.

b) When $\ell(y, g(x)) = (y - g(x))^2$, applying a) we get

$$g^*(x) \in \underset{a \in \mathbb{R}}{\arg\min} \underbrace{\mathbb{E}\left[(Y-a)^2 \mid X=x\right]}_{F(a)}$$

We have $F(a) = \mathbb{E}\left[Y^2 \mid X=x\right] - 2a\,\mathbb{E}\left[Y \mid X=x\right] + a^2$.

The minimum of this function is atteined when

$$a = \mathbb{E}\left[Y \mid X=x\right].$$

c) For $\ell(y, a) = \mathbb{1}(y \neq a)$ we have

$$\begin{aligned}
\underset{a \in \{0,1\}}{\arg\min} \quad & \mathbb{E}\left[\mathbb{1}(Y \neq a) \mid X=x\right] \\
&= \underset{a \in \{0,1\}}{\arg\min} \ \mathbb{P}(Y \neq a \mid X=x) \\
&= \underset{a \in \{0,1\}}{\arg\max} \ \mathbb{P}(Y = a \mid X=x) \\
&= \begin{cases} 1, & \text{if } \mathbb{P}(Y=1 \mid X=x) > 1/2 \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

III Empirical risk minimization

$(X_i, Y_i) \overset{iid}{\sim} P \qquad g: \mathcal{X} \to \mathcal{Y} \quad R_P(g) = \mathbb{E}\left[\ell(Y, g(X))\right].$

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

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

is a good approximation of $R_P(g)$. Indeed, according to the central limit theorem

$$\widehat{R}_n(g) - R_P(g) \approx \frac{\xi(g)}{\sqrt{n}}$$

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

$$\sup_{g \in \mathcal{G}} \left(\widehat{R}_n(g) - R_P(g) \right)$$

does not necessarily go to 0 when $n \longrightarrow +\infty$.

<u>DEF</u> Given a set of candidate prediction functions, $\mathcal{G}$, we call empirical risk minimizer (ERM) the function

$$\widehat{g}_n \in \arg\min_{g \in \mathcal{G}} \widehat{R}_n(g).$$

The choice of the set $\mathcal{G}$ is of central importance. This is clear from the following decomposition:

$$R_P(\widehat{g}_n) - R_P(g^*) = \underbrace{R_P(\widehat{g}_n) - R_P(g^*_{\mathcal{G}})}_{T_1} + \underbrace{R_P(g^*_{\mathcal{G}}) - R_P(g^*)}_{T_2}$$

where $g^*_{\mathcal{G}} \in \arg\min_{g \in \mathcal{G}} R_P(g)$.

It is clear that both T_1 and T_2 are $\geqslant 0$.

In addition T_1 increases when $\mathcal{G}$ becomes larger, whereas T_2 decreases when $\mathcal{G}$ 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 $\widehat{g}_n$ **under**fits. When T_2 is too small and T_1 too large, then $\widehat{g}_n$ overfits.