Statistical Learning Theory Notes

Nong Minh Hieu[1]

[1] School of Physical and Mathematical Sciences, Nanyang Technological University (NTU - Singapore)

Contents

1 Probability settings

1.1 Classification problem

Definition 1.1 (Classifier (h)).
*In **classification problems**, we consider pairs (x, y) where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Where:*

- *$\mathcal{X}$ is the space of **feature vectors**.*
- *$\mathcal{Y}$ is the space of **labels**.*

A classifier is a function $h : \mathcal{X} \to \mathcal{Y}$ which aims to assign correct labels to given feature vectors.

Remark : The key assumptions of classification problems are:

- There exists a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$.
- The pairs (x, y) (observed data) are random samples of the random variables pair (X, Y) which has the distribution P_{XY}.

Definition 1.2 (Decomposition of P_{XY}).
We can decompose P_{XY} in either of the following two ways:

$$P_{XY} = P_{X|Y} P_Y$$
$$P_{XY} = P_{Y|X} P_X$$

Which can be understood as two possible ways to generate the pairs (x, y) from the joint distribution P_{XY}.

- *The first way is to generate a random label $y \sim P_Y$. Then, generate the feature vector corresponding to that label $x \sim P_{X|Y=y}$.*
- *The second way is to generate a random vector $x \sim P_X$. Then, generate the label corresponding to that feature vector $y \sim P_{Y|X=x}$.*

Proposition 1.1: Law of total expectation

Given $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The **law of total expectation** states that:

$$\begin{aligned} \mathbb{E}_{XY}\Big[\phi(X, Y)\Big] &= \mathbb{E}_Y\Big[\mathbb{E}_{X|Y}[\phi(X, Y)]\Big] \\ &= \mathbb{E}_X\Big[\mathbb{E}_{Y|X}[\phi(X, Y)]\Big] \end{aligned}$$

Similar to how P_{XY} is decomposed, law of total expectation describes two way of taking the average value:

- Loop through the labels and take average over the feature vectors corresponding to each label.
- Loop through the feature vectors and take average over the labels corresponding to each vector.

Proof (Proposition 1.1).

We have:

$$\begin{aligned}
\mathbb{E}_{XY}\Big[\phi(X,Y)\Big] &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(x,y)P_{XY}(x,y)dydx \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(x,y)P_X(x)P_{Y|X}(y|x)dydx \\
&= \int_{\mathcal{X}} P_X(x)\int_{\mathcal{Y}} \phi(x,y)P_{Y|X}(y|x)dydx \\
&= \int_{\mathcal{X}} P_X(x)\mathbb{E}_{Y|X=x}\Big[\phi(X,Y)\Big]dx \\
&= \mathbb{E}_X\Big[\mathbb{E}_{Y|X}\Big[\phi(X,Y)\Big]\Big]
\end{aligned}$$

Applying the same technique, we have $\mathbb{E}_{XY}\Big[\phi(X,Y)\Big] = \mathbb{E}_Y\Big[\mathbb{E}_{X|Y}[\phi(X,Y)]\Big]$. $\square$.

Remark : Usually, the label space is discrete and finite, meaning $\mathcal{Y} = \{0, 1, 2, \ldots, m\}$ for some $m < \infty$. Hence, the expectations over Y can be written as discrete sums:

$$\begin{aligned}
\mathbb{E}_{XY}\Big[\phi(X,Y)\Big] = \mathbb{E}_Y\Big[\mathbb{E}_{X|Y}[\phi(X,Y)]\Big] &= \sum_{y\in\mathcal{Y}} \mathbb{E}_{X|Y=y}[\phi(X,Y)] \\
= \mathbb{E}_X\Big[\mathbb{E}_{Y|X}[\phi(X,Y)]\Big] &= \mathbb{E}_X\left[\sum_{y\in\mathcal{Y}} \mathbb{E}_{Y=y|X}[\phi(X,Y)]\right]
\end{aligned}$$

Definition 1.3 (Hypothesis space ($\mathcal{H}$)).

The hypothesis space is a collection (family) of classifiers $h : \mathcal{X} \to \mathcal{Y}$ that have some common properties:

$$\mathcal{H} = \Big\{h : \mathcal{X} \to \mathcal{Y}\Big|\textit{some common properties}\Big\}$$

For example, let $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = (0,1)$. In logistic regression, we assume the classifiers to be logit functions:

$$\mathcal{H}_{logit} = \Big\{h : \mathbb{R}^d \to (0,1)\Big|h(x) = logit(\beta x) = \frac{1}{1+e^{-\beta x}}, \beta \in \mathbb{R}^{1\times d}\Big\}$$

Definition 1.4 (Learning algorithm ($\mathcal{L}_n$)).

*To learn a classifier $h : \mathcal{X} \to \mathcal{Y}$, suppose that we have access to a training dataset of n data pairs $\{(X_k, Y_k)\}_{k=1}^n$ which are assumed to be **i.i.d sampled from** P_{XY}. The domain of the training data is then $(\mathcal{X}\times\mathcal{Y})^n$. A **learning algorithm**, denoted as $\mathcal{L}_n$ is a function/procedure that derives a classifier $\hat{h}_n : \mathcal{X} \to \mathcal{Y}$ from the training data.*

$$\begin{aligned}
&\mathcal{L}_n : (\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H} \\
&\hat{h}_n = \mathcal{L}_n((X_1, Y_1), \ldots, (X_n, Y_n))
\end{aligned}$$

1.2 Goal of classification

Definition 1.5 (Risk ($R(h)$)).
The ***risk*** *of a classifier is defined as followed:*

$$R(h) = P(h(X) \neq Y) = \mathbb{E}[\mathbf{1}_{\{h(X)\neq Y\}}]$$

Where (X, Y) *are independent of the training data.*

Definition 1.6 (Bayes Risk (R^*)).
The ***Bayes risk*** *is the infimum of the risk taken over all* $h : \mathcal{X} \to \mathcal{Y}$*, not just for* $h \in \mathcal{H}$*:*

$$R^* = \inf_{h:\mathcal{X}\to\mathcal{Y}} R(h)$$

Definition 1.7 (Consistency of learning algorithms).
A learning algorithm $\mathcal{L}_n$ *is called:*

- ***Weakly consistent*** *if* $R(\hat{h}_n) \xrightarrow{p} R^*$*:*

$$\lim_{n\to\infty} P(R(\hat{h}_n) \leq r) = P(R^* \leq r),\ \forall r \geq 0$$

- ***Strongly consistent*** *if* $R(\hat{h}_n) \xrightarrow{a.s} R^*$*:*

$$P\Big(\lim_{n\to\infty} \Big|R(\hat{h}_n) - R^*\Big| \geq \epsilon\Big) = 0,\ \forall \epsilon > 0$$

- ***Universally weakly/strongly consistent*** *if* $\mathcal{L}_n$ *is weakly/strongly consistent for all* P_{XY}*. Meaning, consistency holds without any assumption about* P_{XY}*.*

2 Bayes classifier

2.1 Properties of Bayes Risk

Overview : Recall that the Bayes classifier is the one with minimum risk and the corresponding risk is called the Bayes Risk. For $\mathcal{Y} = \{0, 1\}$ and defined:

$$\eta(x) = P(Y = 1|X = x)$$

Define the following classifier:

$$h^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ \\ 0 & \text{otherwise} \end{cases}$$

Theorem 2.1: Properties of Bayes classifier

The following properties hold for the Bayes classifier with $\mathcal{Y} = \{0, 1\}$ (Binary classification):

- (i) $R(h^*) = \inf_{h:\mathcal{X}\to\mathcal{Y}}\{R(h)\} = R^*$.
- (ii) $\underbrace{R(h) - R^*}_{\text{Excess risk}} = 2\mathbb{E}_X\left[\left|\eta(x) - \frac{1}{2}\right|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\right]$.
- (iii) $R^* = \mathbb{E}\Big[\min(\eta(X), 1 - \eta(x))\Big]$.

Proof (Theorem 2.1).

Proving each point:

(i) $R(h^*) = \inf_{h:\mathcal{X}\to\mathcal{Y}}\{R(h)\} = R^*$.
For all $h : \mathcal{X} \to \mathcal{Y}$*, we have:*

$$\begin{aligned} R(h) &= \mathbb{E}_{XY}\Big[\mathbf{1}_{\{h(X)\neq Y\}}\Big] \\ &= \mathbb{E}_{x\sim X}\left[\mathbb{E}_{Y|X=x}\Big[\mathbf{1}_{\{Y\neq h(x)\}}\Big]\right] \\ &= \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}}\mathbf{1}_{\{y\neq h(x)\}}\right] \\ &= \mathbb{E}_{x\sim X}\Big[\eta(x)\mathbf{1}_{\{h(x)=0\}} + (1-\eta(x))\mathbf{1}_{\{h(x)=1\}}\Big] \end{aligned}$$

Since the two events $\{h(x) = 1\}$ *and* $\{h(x) = 0\}$ *are mutually exclusive,* $R(h)$ *is the smallest when we set* $h(x) = 1$ *when* $\eta(x) \geq 1 - \eta(x) \implies \eta(x) \geq \frac{1}{2}$*. Therefore, we have:*

$$h^*(x) = \begin{cases} 1 & \textit{if } \eta(x) \geq \frac{1}{2} \\ \\ 0 & \textit{otherwise} \end{cases}$$

(ii) $\underbrace{R(h) - R^*}_{\textit{Excess risk}} = 2\mathbb{E}_X\left[\left|\eta(x) - \frac{1}{2}\right|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\right]$.

We have:

$$\begin{aligned}
R(h) - R^* &= \mathbb{E}_{x\sim X}\left[\mathbb{E}_{Y|X=x}\Big[\mathbf{1}_{\{Y\neq h(x)\}}\Big]\right] - \mathbb{E}_{x\sim X}\left[\mathbb{E}_{Y|X=x}\Big[\mathbf{1}_{\{Y\neq h^*(x)\}}\Big]\right] \\
&= \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}} \mathbf{1}_{\{y\neq h(x)\}} P(Y=y|X=x)\right] - \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}} \mathbf{1}_{\{y\neq h^*(x)\}} P(Y=y|X=x)\right] \\
&= \mathbb{E}_{x\sim X}\left[\eta(x)\Big(\mathbf{1}_{\{h(x)=0\}} - \mathbf{1}_{\{h^*(x)=0\}}\Big) + (1-\eta(x))\Big(\mathbf{1}_{\{h(x)=1\}} - \mathbf{1}_{\{h^*(x)=1\}}\Big)\right] \\
&= \mathbb{E}_{x\sim X}\Bigg[\eta(x)\Big(\mathbf{1}_{\{h(x)\neq h^*(x),h(x)=0\}} - \mathbf{1}_{\{h(x)\neq h^*(x),h(x)=1\}}\Big) \\
&\qquad + (1-\eta(x))\Big(\mathbf{1}_{\{h(x)\neq h^*(x),h(x)=1\}} - \mathbf{1}_{\{h(x)\neq h^*(x),h(x)=0\}}\Big)\Bigg] \\
&= \mathbb{E}_{x\sim X}\left[(2\eta(x)-1)\mathbf{1}_{\{h(x)\neq h^*(x),h(x)=0\}} + (1-2\eta(x))\mathbf{1}_{\{h(x)\neq h^*(x),h(x)=1\}}\right] \\
&= \mathbb{E}_{x\sim X}\left[\Big|2\eta(x)-1\Big|\mathbf{1}_{\{h(x)\neq h^*(x)\}}\right] \\
&= 2\mathbb{E}_X\left[\left|\eta(X)-\frac{1}{2}\right|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\right]
\end{aligned}$$

(iii) $R^* = \mathbb{E}\Big[\min(\eta(X), 1-\eta(x))\Big]$**.**
From (i) we have:

$$\begin{aligned}
R(h^*) &= \mathbb{E}_{x\sim X}\Big[\eta(x)\mathbf{1}_{\{h^*(x)=0\}} + (1-\eta(x))\mathbf{1}_{\{h^*(x)=1\}}\Big] \\
&= \mathbb{E}_X\Big[\min(\eta(X), 1-\eta(x))\Big]
\end{aligned}$$

$\square$.

Theorem 2.2: Properties of Bayes classifier (Multi-class)

For multi-class classification with more than two labels : $\mathcal{Y} = \{1, 2, \ldots, M\}$, the Bayes classifier is defined as followed:

$$h^*(x) = \arg\max_{y\in\mathcal{Y}} \Big\{\eta_y(x)\Big\}$$
$$\text{Where : } \eta_y(x) = P(Y = y|X = x)$$

The following properties hold for the Bayes classifier with $\mathcal{Y} = \{1, 2, \ldots, M\}$ (Multi-class classification):

- (i) **Bayes Risk** R^* :

$$R^* = \mathbb{E}_{x\sim X}\Big[1 - \max_{y\in\mathcal{Y}}\Big\{\eta_y(x)\Big\}\Big] = \mathbb{E}_{x\sim X}\Big[\min_{y\in\mathcal{Y}} \overline{\eta_y}(x)\Big]$$

- (ii) **Excess Risk** $R(h) - R^*$:

$$R(h) - R^* = \mathbb{E}_X\Big[\Big(\eta_{y_x^*}(x) - \eta_{y_x}(x)\Big)\mathbf{1}_{\{h(x)\neq h^*(x)\}}\Big]$$

Where $y_x = h(x)$ is the prediction made by an arbitrary classifier $h : \mathcal{X} \to \mathcal{Y}$ and $y_x^* = h^*(x)$ is the prediction made by the Bayes classifier.

Proof (Theorem 2.2)**.**
(The proof of this theorem has been included in the solution of Exercise 2.1). □.

2.2 Likelihood Ratio Test

Overview : Define $\pi_1 = P(Y = 1)$ and $\pi_0 = P(Y = 0)$ be the prior probabilities. Let $p_1(x) = P(X = x|Y = 1)$ and $p_0(x) = P(X = x|Y = 0)$ be the class-conditional densities. Note that we have:

$$\begin{aligned}
\eta(x) &= P(Y = 1|X = x) \\
&= \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 1)P(Y = 1) + P(X = x|Y = 0)P(Y = 0)} \\
&= \frac{\pi_1 p_1(x)}{\pi_1 p_1(x) + \pi_0 p_0(x)} \\
&= \frac{1}{1 + \frac{\pi_0 p_0(x)}{\pi_1 p_1(x)}}
\end{aligned}$$

Hence, we have:

$$\begin{aligned}
\eta(x) \geq \frac{1}{2} &\iff \frac{\pi_0 p_0(x)}{\pi_1 p_1(x)} \\
&\iff \frac{p_1(x)}{p_0(x)} \geq \frac{\pi_0}{\pi_1}
\end{aligned}$$

Proposition 2.1: Likelihood ratio test

The Bayes classifier h^* can be re-defined as followed:

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} \geq \frac{\pi_0}{\pi_1} \\ \\ 0 & \text{otherwise} \end{cases}$$

The fraction $\frac{p_1(x)}{p_0(x)}$ is called the **likelihood ratio**.

2.3 Plug-in classifier

Definition 2.1 (Plug-in classifier).

*A **plug-in classifier** is based on an estimate of $\eta(x)$. This estimate is then plugged into the definition of the Bayes classifier. Suppose that $\widehat{\eta_n}$ is an estimate of η based on n training samples $\{(X_i, Y_i)\}_{i=1}^n$. We define $\widehat{h_n}$ as:*

$$\widehat{h_n} = \begin{cases} 1 & \textit{if } \widehat{\eta_n}(x) \geq \frac{1}{2} \\ \\ 0 & \textit{otherwise} \end{cases}$$

Corollary 2.1: Excess risk of plug-in classifier

We have the following upper-bound for the excess risk of the plug-in classifier:

$$R(\widehat{h_n}) - R^* \leq 2\mathbb{E}_X\left[\left|\eta(X) - \widehat{\eta_n}(X)\right|\right]$$

Proof (Corollary 2.1).

From theorem 2.1, we have:

$$R(\widehat{h_n}) - R^* = 2\mathbb{E}_X\left[\left|\eta(X) - \frac{1}{2}\right| \mathbf{1}_{\{\widehat{h_n}(X) \neq h^*(X)\}}\right]$$

The indicator term will be non-zero in the above equality if one of the following cases occurs:

$$\begin{cases} \widehat{h_n}(X) = 1, h^*(X) = 0 \\ \\ \widehat{h_n}(X) = 0, h^*(X) = 1 \end{cases} \implies \begin{cases} \widehat{\eta_n}(X) \geq \frac{1}{2}, \eta(X) < \frac{1}{2} \\ \\ \widehat{\eta_n}(X) < \frac{1}{2}, \eta(X) \geq \frac{1}{2} \end{cases}$$

Case 1 : $\widehat{\eta_n}(X) \geq \frac{1}{2}, \eta(X) < \frac{1}{2}$

We have:

$$\begin{aligned} \eta(X) - \widehat{\eta_n}(X) &\leq \eta(X) - \frac{1}{2} \quad (\textit{Both sides negative}) \\ \implies \left|\eta(X) - \widehat{\eta_n}(X)\right| &\geq \left|\eta(X) - \frac{1}{2}\right| \end{aligned}$$

Case 2 : $\widehat{\eta_n}(X) < \frac{1}{2}, \eta(X) \geq \frac{1}{2}$

We have:

$$\widehat{\eta_n}(X) - \eta(X) \geq \widehat{\eta_n}(X) - \frac{1}{2} \geq \eta(X) - \frac{1}{2} \quad \textit{(All positive)}$$

Therefore, we have:

$$\left|\eta(X) - \widehat{\eta_n}(X)\right| \geq \left|\eta(X) - \frac{1}{2}\right|$$

For both cases, we have the same $\left|\eta(X) - \widehat{\eta_n}(X)\right| \geq \left|\eta(X) - \frac{1}{2}\right|$ *inequality. Therefore, we have:*

$$R(\widehat{h_n}) - R^* \leq 2\mathbb{E}_X\left[\left|\eta(X) - \widehat{\eta_n}(X)\right|\right]$$

$\square$.

2.4 End of chapter exercises

Exercise 2.1

Extend theorem 2.1 to the multi-class classification case where $\mathcal{Y} = \{1, 2, \ldots, M\}$. In other words, prove theorem 2.2.

Solution (Exercise 2.1).

We re-define the Bayes classifier h^* as followed:

$$h^*(x) = \arg\max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\},$$
$$\eta_y(x) = P(Y = y | X = x)$$

We have:

$$\sum_{y \in \mathcal{Y}} \eta_y(x) = 1, \ \forall x \in \mathcal{X}$$

(i) *Calculate Bayes risk* R^*

For any classifier $h : \mathcal{X} \to \mathcal{Y}$, we have:

$$R(h) = \mathbb{E}_{x \sim X} \left[\sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x) \right]$$

Letting $\hat{y}_x = h(x)$ being h's prediction for a given feature vector $x \in \mathcal{X}$, we have:

$$R(h) = \mathbb{E}_{x \sim X} \left[\sum_{y \in \mathcal{Y}; y \neq \hat{y}_x} \eta_y(x) \right] = \mathbb{E}_{x \sim X} \left[1 - \eta_{\hat{y}_x}(x) \right]$$

In order to minimize $R(h)$, we need $\eta_{\hat{y}_x}(x)$ to be maxmized for all $x \in \mathcal{X}$. Hence, we have:

$$R^* = \mathbb{E}_{x \sim X} \left[1 - \max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\} \right]$$

Therefore, we have $h^*(x) = \arg\max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\}$ is the Bayes classifier and the Bayes risk $R^* = \mathbb{E}_{x \sim X} \left[1 - \max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\} \right]$.

(ii) *Calculate excess risk* $R(h) - R^*$

For any $h : \mathcal{X} \to \mathcal{Y}$, we have:

$$\begin{aligned} R(h) - R^* &= \mathbb{E}_{x \sim X} \left[\sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x) \right] - \mathbb{E}_{x \sim X} \left[1 - \max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\} \right] \\ &= \mathbb{E}_{x \sim X} \left[\sum_{y \in \mathcal{Y}} \mathbf{1}_{\{h(x) \neq y\}} \eta_y(x) + \max_{y \in \mathcal{Y}} \Big\{ \eta_y(x) \Big\} - 1 \right] \end{aligned}$$

Denote $h^*(x) = y_x^*$ and $h(x) = y_x$. When $h(x) = h^*(x) = y_x^*$, we have:

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} \mathbf{1}_{\{h(x)\neq y\}}\eta_y(x) + \max_{y\in\mathcal{Y}}\Big\{\eta_y(x)\Big\} &= \sum_{y\in\mathcal{Y};y\neq y_x} \eta_y(x) + \eta_{y_x^*}(x) \\
&= \sum_{y\in\mathcal{Y};y\neq y_x^*} \eta_y(x) + \eta_{y_x^*}(x) \\
&= \sum_{y\in\mathcal{Y}} \eta_y(x) = 1 \\
\implies \sum_{y\in\mathcal{Y}} \mathbf{1}_{\{h(x)\neq y\}}\eta_y(x) + \max_{y\in\mathcal{Y}}\Big\{\eta_y(x)\Big\} - 1 = 0
\end{aligned}$$

When $h(x) \neq h^*(x)$, we have:

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} \mathbf{1}_{\{h(x)\neq y\}}\eta_y(x) + \max_{y\in\mathcal{Y}}\Big\{\eta_y(x)\Big\} - 1 &= \sum_{y\in\mathcal{Y};y\neq y_x} \eta_y(x) + \eta_{y_x^*}(x) - 1 \\
&= 2\eta_{y_x^*}(x) - 1 + \sum_{y\in\mathcal{Y}\backslash\{y_x,y_x^*\}} \eta_y(x) \\
&= 2\eta_{y_x^*}(x) - \Big(\eta_{y_x}(x) + \eta_{y_x^*}(x)\Big) \\
&= \eta_{y_x^*}(x) - \eta_{y_x}(x).
\end{aligned}$$

Therefore, we can re-write the excess risk by multiplying the entire integrand with the indicator function $\mathbf{1}_{\{h(x)\neq h^*(x)\}}$ as followed:

$$R(h) - R^* = \mathbb{E}_{x\sim X}\left[\Big(\eta_{y_x^*}(x) - \eta_{y_x}(x)\Big)\mathbf{1}_{\{h(x)\neq h^*(x)\}}\right]$$

(iii) ***Simpler form of Bayes risk***
From (i) we have:

$$R^* = \mathbb{E}_X\left[1 - \max_{y\in\mathcal{Y}}\Big\{\eta_y(x)\Big\}\right] = \mathbb{E}_X\left[\min_{y\in\mathcal{Y}}\Big\{\overline{\eta_y}(x)\Big\}\right]$$

Where $\overline{\eta_y}(x) = P(Y \neq y|X = x)$.

$\square$.

Exercise 2.2

Define the α**-cost-sensitive risk** of a classifier $h : \mathcal{X} \to \mathcal{Y}$ as followed:

$$R_\alpha(h) = \mathbb{E}_{XY}\Big[(1-\alpha)\mathbf{1}_{\{Y=1,h(X)=0\}} + \alpha\mathbf{1}_{\{Y=0,h(X)=1\}}\Big]$$

Define the Bayes classifier and prove and analogue of theorem 2.1.

Solution (Exercise 2.2).
Using the law of total expectation, we have:

$$\begin{aligned}
R_\alpha(h) &= \mathbb{E}_{x\sim X}\left[\sum_{y\in\{0,1\}}\Big[(1-\alpha)\mathbf{1}_{\{y=1,h(x)=0\}} + \alpha\mathbf{1}_{\{y=0,h(x)=1\}}\Big]P(Y=y|X=x)\right] \\
&= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\mathbf{1}_{\{h(x)=0\}} + \alpha(1-\eta(x))\mathbf{1}_{\{h(x)=1\}}\Big]
\end{aligned}$$

Since $\mathbf{1}_{\{h(x)=0\}}$ and $\mathbf{1}_{\{h(x)=1\}}$ are mutually exclusive, in order for $R_\alpha(h)$ to be minimize, we define the following Bayes classifier:

$$h^*(x) = \begin{cases} 1 & \textit{if } \alpha(1-\eta(x)) \le (1-\alpha)\eta(x) \\ \\ 0 & \textit{otherwise} \end{cases} = \begin{cases} 1 & \textit{if } \eta(x) \ge \alpha \\ \\ 0 & \textit{otherwise} \end{cases}$$

We can also derive a likelihood-ratio test version of the Bayes classifier, we have:

$$\begin{aligned} \eta(x) \ge \alpha &\implies \frac{1}{1+\frac{\pi_0 p_0(x)}{\pi_1 p_1(x)}} \ge \alpha \\ &\implies 1 + \frac{\pi_0 \cdot p_0(x)}{\pi_1 \cdot p_1(x)} \le \frac{1}{\alpha} \\ &\implies \frac{p_1(x)}{p_0(x)} \ge \frac{\alpha}{1-\alpha} \cdot \frac{\pi_0}{\pi_1} \end{aligned}$$

Hence, we can rewrite the Bayes classifier as followed:

$$h^*(x) = \begin{cases} 1 & \textit{if } \frac{p_1(x)}{p_0(x)} \ge \frac{\alpha}{1-\alpha} \cdot \frac{\pi_0}{\pi_1} \\ \\ 0 & \textit{otherwise} \end{cases}$$

(**i**) ***Bayes Risk*** R^*_α
We have:

$$\begin{aligned} R^*_\alpha &= R_\alpha(h^*) \\ &= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\mathbf{1}_{\{h^*(x)=0\}} + \alpha(1-\eta(x))\mathbf{1}_{\{h^*(x)=1\}}\Big] \\ &= \mathbb{E}_X\Big[\min(\alpha(1-\eta(X)), (1-\alpha)\eta(X))\Big] \end{aligned}$$

(**ii**) ***Excess Risk*** $R_\alpha(h) - R^*_\alpha$
For an arbitrary $h : \mathcal{X} \to \mathcal{Y}$, we have:

$$\begin{aligned} R_\alpha(h) - R^*_\alpha &= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\Big(\mathbf{1}_{\{h(x)=0\}} - \mathbf{1}_{\{h^*(x)=0\}}\Big) + \alpha(1-\eta(x))\Big(\mathbf{1}_{\{h(x)=1\}} - \mathbf{1}_{\{h^*(x)=1\}}\Big)\Big] \\ &= \mathbb{E}_{x\sim X}\Big[(1-\alpha)\eta(x)\Big(\mathbf{1}_{\{h(x)=0,h^*(x)=1\}} - \mathbf{1}_{\{h(x)=1,h^*(x)=0\}}\Big) \\ &\quad + \alpha(1-\eta(x))\Big(\mathbf{1}_{\{h(x)=1,h^*(x)=0\}} - \mathbf{1}_{\{h(x)=0,h^*(x)=1\}}\Big)\Big] \\ &= \mathbb{E}_{x\sim X}\Big[\mathbf{1}_{\{h(x)=0,h^*(x)=1\}}(\eta(x)-\alpha) + \mathbf{1}_{\{h(x)=1,h^*(x)=0\}}(\alpha-\eta(x))\Big] \\ &= \mathbb{E}_X\Big[\Big|\eta(X)-\alpha\Big|\mathbf{1}_{\{h(X)\neq h^*(X)\}}\Big] \end{aligned}$$

$\square$.

3 Hoeffding's inequality

3.1 Markov's Inequality

Proposition 3.1: Markov's Inequality

Let U be a non-negative random variable on $\mathbb{R}$, then for all $t > 0$, we have:

$$P(U \geq t) \leq \frac{1}{t}\mathbb{E}[U]$$

Proof (Proposition 3.1).

We have:

$$\begin{aligned}
tP(U \geq t) &= t\mathbb{E}\Big[\mathbf{1}_{\{U \geq t\}}\Big] \\
&= t\int_0^\infty \mathbf{1}_{\{x \geq t\}} f_U(x)dx \\
&= t\int_t^\infty f_U(x)dx \\
&\leq \int_t^\infty x f_U(x)dx \\
&\leq \int_0^\infty x f_U(x)dx = \mathbb{E}[U] \\
\implies P(U \geq t) &\leq \frac{1}{t}\mathbb{E}[U]
\end{aligned}$$

□.

Corollary 3.1: Chebyshev's Inequality

Let Z be a random variable on $\mathbb{R}$ with mean μ and variance σ^2, we have:

$$P\Big(\Big|Z - \mu\Big| \geq t\Big) \leq \frac{\sigma^2}{t^2}$$

Proof (Corollary 3.1).

Using Markov's inequality, we have:

$$\begin{aligned}
P\Big(\Big|Z - \mu\Big| \geq t\Big) &= P\Big(\Big|Z - \mu\Big|^2 \geq t^2\Big) \\
&\leq \frac{\mathbb{E}\Big[\Big|Z - \mu\Big|^2\Big]}{t^2} = \frac{\sigma^2}{t^2}
\end{aligned}$$

□.

Corollary 3.2: Chernoff's bounding method

Let Z be a random variable on $\mathbb{E}$, for any $t > 0$, we have:

$$P(Z \geq t) \leq \inf_{s>0} e^{-st} M_Z(s)$$

Proof (Corollary 3.2).

We have:

$$
\begin{aligned}
P(Z \geq t) &= P(sZ \geq st), \quad (t > 0) \\
&= P(e^{sZ} \geq e^{st}) \\
&\leq \frac{\mathbb{E}\left[e^{sZ}\right]}{e^{st}} = e^{-st} M_Z(s) \quad \textit{(Markov's inequality)}
\end{aligned}
$$

Since the above inequality holds for all $s > 0$, we can just take the infimum to obtain the tightest bound. Hence, we have:

$$P(Z \geq t) \leq \inf_{s>0} e^{-st} M_Z(s)$$

$\square$.

3.2 Hoeffding's Inequality

Before diving into Hoeffding's inequality, we need to go through the following lemma (whose proof will not be included) that will help us prove the Hoeffding's inequality:

Lemma 3.1: Hoeffding's lemma

Let V be a random variable on $\mathbb{R}$ with $\mathbb{E}[V] = 0$ and suppose that $a \leq V \leq b$ with probability one. We have:

$$\mathbb{E}\left[e^{sV}\right] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Proof (Lemma 3.1).

(The proof for this lemma can be found here [3]). $\square$.

Theorem 3.1: Hoeffding's Inequality

Let $Z_1, Z_2, \dots, Z_n$ be independent random variables on $\mathbb{R}$ such that $a_i \leq Z_i \leq b_i$ with probability one for all $1 \leq i \leq n$. Let $S_n = \sum_{i=1}^n Z_i$. We have:

$$P\left(\left|S_n - \mathbb{E}[S_n]\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad \forall t > 0$$

Proof (Theorem 3.1).

Using the Chernoff's bounds, we have:

$$
\begin{aligned}
P\Big(\Big|S_n - \mathbb{E}[S_n]\Big| \geq t\Big) &\leq \inf_{s>0} e^{-st} M_{S_n-\mathbb{E}[S_n]}(s) \\
&= \inf_{s>0} e^{-st}\mathbb{E}\Big[e^{s(S_n-\mathbb{E}[S_n])}\Big] \\
&= \inf_{s>0} e^{-st}\mathbb{E}\left[\exp\left(s\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right)\right] \\
&= \inf_{s>0} e^{-st}\mathbb{E}\left[\prod_{i=1}^{n}\exp\Big(s(Z_i - \mathbb{E}[Z_i])\Big)\right] \\
&= \inf_{s>0} e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[\exp\Big(s(Z_i - \mathbb{E}[Z_i])\Big)\right] \quad \textit{(Since all } Z_i - \mathbb{E}[Z_i] \textit{ are independent)} \\
&\leq \inf_{s>0} e^{-st}\prod_{i=1}^{n}\exp\left(\frac{s^2(b_i-a_i)^2}{8}\right) \quad \textit{(By Hoeffding's lemma)} \\
&= \inf_{s>0} \exp\left(-st + \sum_{i=1}^{n}\frac{s^2(b_i-a_i)^2}{8}\right)
\end{aligned}
$$

In order for the above to be minimized, we differentiate the term inside the exponential and set the derivative to 0 *to find the optimal* $s > 0$*. We have:*

$$
-t + s\sum_{i=1}^{n}\frac{(b_i-a_i)^2}{4} = 0 \implies s = \frac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}
$$

Letting $c = \sum_{i=1}^{n}(b_i - a_i)^2$*, we now can derive the tightest Chernoff's bound as followed:*

$$
\begin{aligned}
P\Big(\Big|S_n - \mathbb{E}[S_n]\Big| \geq t\Big) \leq \exp\left(-\frac{4t^2}{c} + \frac{16t^2}{c^2}\cdot\frac{c}{8}\right) &= \exp\left(-\frac{2t^2}{c}\right) \\
&= \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)
\end{aligned}
$$

$\square$.

3.3 Convergence of Empirical Risk

Definition 3.1 (Empirical Risk)**.**
Suppose we are given training data $\Big\{(X_i, Y_i)_{i=1}^{n}\Big\}$ *such that each pair* $(X_i, Y_i) \sim P_{XY}$ *are independently identically distributed. Let* $h : \mathcal{X} \to \mathcal{Y}$ *be a classifier. We define the* ***empirical risk*** *to be:*

$$
\widehat{R_n}(h) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{h(X_i)\neq Y_i\}}
$$

Note that $\mathbb{E}[\widehat{R_n}(h)] = R(h)$ *and* $n\widehat{R_n}(h) \sim Binomial(n, R(h))$*. In the following corollary of the Hoeffding's inequality, we will answer the question* ***how close the empirical risk is as an estimate of true risk*** *or* ***how fast the empirical risk converges to the true risk****.*

Corollary 3.3: Convergence of Empirical Risk

Given training data $\left\{(X_i, Y_i)_{i=1}^n\right\}$ such that each pair $(X_i, Y_i) \sim P_{XY}$ are independently identically distributed. Let $h : \mathcal{X} \to \mathcal{Y}$ be a classifier, we have:

$$P\left(\left|\widehat{R_n}(h) - R(h)\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}, \quad \epsilon > 0$$

Proof (Corollary 3.3).

For all $1 \leq i \leq n$*, we have* $\mathbf{1}_{\{h(X_i) \neq Y_i\}} \in \{0, 1\}$*. Hence, with probability one,* $0 \leq \mathbf{1}_{\{h(X_i) \neq Y_i\}} \leq 1$ *and* $b_i = 1, a_i = 0$ *for all* $1 \leq i \leq n$*.*

Using the Hoeffding's inequality, we have:

$$\begin{aligned}
P\left(\left|\widehat{R_n}(h) - R(h)\right| \geq \epsilon\right) &= P\left(\left|\widehat{R_n}(h) - \mathbb{E}[\widehat{R_n}(h)]\right| \geq \epsilon\right) \\
&= P\left(\left|n\widehat{R_n}(h) - \mathbb{E}[n\widehat{R_n}(h)]\right| \geq n\epsilon\right) \\
&\leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \textit{(Hoeffding's inequality)} \\
&= e^{-2n\epsilon^2}
\end{aligned}$$

$\square$.

3.4 KL-divergence & Hypothesis Testing

Set-up (Hypothesis Testing) : Suppose that we have $\mathcal{Y} = \{0, 1\}$ and P_{XY} is a distribution on $\mathcal{X} \times \mathcal{Y}$. Let's assume that:

- The prior probabilities π_y are equal.
- The supports of likelihoods p_0, p_1 are the same.
- $0 < \alpha \leq p_y(x) \leq \beta < \infty$ for all $x \in \mathcal{X}$ such that $p_y(x) > 0$ and for all $y \in \{0, 1\}$.

Now suppose $X_1, \ldots, X_n \sim p_y$ are independently identically distributed where $y \in \{0, 1\}$ is unknown. Can we guess y and how good our guess would be?

Proposition 3.2: KL-divergence hypothesis testing

From the above settings, the optimal classifier is given by the likelihood ratio test:

$$\widehat{h_n}(x) = \begin{cases} 1 & \text{if } \frac{\prod_{i=1}^n p_1(x_i)}{\prod_{i=1}^n p_0(x_i)} \geq \frac{\pi_0}{\pi_1} \quad (=1) \\ \\ 0 & \text{otherwise} \end{cases}$$

Where $x = \Big(x_1, \ldots, x_n\Big)$ is an observation of the random vector $X = \Big(X_1, \ldots, X_n\Big)$. Define the class-specific risk $R_y(h)$ be the risk of misclassification when the true label is $Y = y$:

$$R_y(h) = P(h(X) \neq Y | Y = y)$$

Then, we have:

$$R_0(\widehat{h_n}) \leq e^{-2nD(p_0||p_1)^2/c}, \text{ where } c = 4(\log\beta - \log\alpha)^2$$

Where $D(p_0||p_1)$ is the KL-divergence of p_1 from p_0. We can prove a similar exponentially decaying bound for $R_1(\widehat{h_n})$.

Proof.

Proposition 3.2 We can rewrite the optimal classifier as:

$$\widehat{h_n}(X) = \begin{cases} 1 & \textit{if } \widehat{S_n}(X_1, \ldots, X_n) \geq 0 \\ \\ 0 & \textit{otherwise} \end{cases}$$

Where we have:

$$\begin{aligned} \widehat{S_n}(X_1, \ldots, X_n) &= \log \frac{\prod_{i=1}^n p_1(X_i)}{\prod_{i=1}^n p_0(X_i)} \\ &= \sum_{i=1}^n \log \frac{p_1(X_i)}{p_0(X_i)} \\ &= \sum_{i=1}^n Z_i \quad \left(\textit{Letting } Z_i = \log\frac{p_1(X_i)}{p_0(X_i)}\right) \end{aligned}$$

Since the likelihoods are bounded, we have:

$$a_i = \log\frac{\alpha}{\beta} \leq Z_i \leq \log\frac{\beta}{\alpha} = b_i, \quad 1 \leq i \leq n$$

Now, we have:

$$\begin{aligned} R_0(\widehat{h_n}) &= P(h(X) \neq Y | Y = 0) \\ &= P(\widehat{S_n} \geq 0 | Y = 0) \\ &= P(\widehat{S_n} - \mathbb{E}[S_n | Y = 0] \geq -\mathbb{E}[S_n | Y = 0] | Y = 0) \end{aligned}$$

To calculate the conditional expectation $\mathbb{E}[S_n|Y = 0]$, we have:

$$\begin{aligned} \mathbb{E}[S_n | Y = 0] &= n\mathbb{E}[Z_1 | Y = 0] \\ &= n\int \log\frac{p_1(x)}{p_0(x)} p_0(x) dx \\ &= -n\int \log\frac{p_0(x)}{p_1(x)} p_0(x) dx = -nD(p_0||p_1) \end{aligned}$$

Therefore, we have:

$$\begin{aligned} R_0(\widehat{h_n}) &= P(\widehat{S_n} - \mathbb{E}[S_n|Y=0] \geq nD(p_0||p_1)|Y=0) \\ &\leq \exp\left(- \frac{2n^2 D(p_0||p_1)^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \quad \textit{(Hoeffding's inequality)} \end{aligned}$$

For every $1 \leq i \leq n$*, we have:*

$$\begin{aligned} b_i - a_i &= \log \frac{\beta}{\alpha} - \log \frac{\alpha}{\beta} \\ &= \log \frac{\beta^2}{\alpha^2} = 2 \log \frac{\beta}{\alpha} = 2(\log \beta - \log \alpha) \\ \implies \sum_{i=1}^n (b_i - a_i)^2 &= 4n(\log \beta - \log \alpha)^2 \end{aligned}$$

Finally, we have:

$$R_0(\widehat{h_n}) \leq \exp\left(- \frac{2nD(p_0||p_1)^2}{4(\log \beta - \log \alpha)^2} \right)$$

Similarly, for $R_1(\widehat{h_n})$*, we have:*

$$R_1(\widehat{h_n}) \leq \exp\left(- \frac{2nD(p_1||p_0)^2}{4(\log \beta - \log \alpha)^2} \right)$$

$\square$.

A List of Definitions

B Important Theorems

C Important Corollaries

D Important Propositions

E References

References

[1] Rick Durrett. *Probability: Theory and Examples*. 4th. USA: Cambridge University Press, 2010. ISBN: 0521765390.

[2] Erhan undefinedinlar. *Probability and Stochastics*. Springer New York, 2011. ISBN: 9780387878591. DOI: 10.1007/978-0-387-87859-1. URL: http://dx.doi.org/10.1007/978-0-387-87859-1.

[3] Wikipedia. *Hoeffding's lemma — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Hoeffding's%20lemma&oldid=1114715065. [Online; accessed 04-January-2024]. 2024.

[4] Wikipedia. *Vitali set — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Vitali%20set&oldid=1187241923. [Online; accessed 24-December-2023]. 2023.