Statistical Learning Theory - Exercise #1 (Chapters 1,2).

1 Problem 1.

↳ Theorem 1.1: Let $\mathcal{X}$ be the space of features and $\mathcal{Y} = \{0,1\}$ be the labels space for a binary classification problem. Define the Bayes Classifier h^* as followed:

$$h^*(X) = \begin{cases} 1 & \text{if } p(X) \geq \frac{1}{2} \\ 0 & \text{Otherwise} \end{cases}$$

Where $p(X) = P(Y=1|X)$. Then, we have the following properties of h^*:

(i) $R(h^*) = R^* \Rightarrow h^*$ is the Bayes Classifier.

(ii) $\forall h: \mathcal{X} \to \mathcal{Y}$:

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

(iii) $R^* = \mathbb{E}_X\left[\min(p(X), 1-p(X))\right]$.

↳ Problem: Extend the above theorem to multi-class classification problem. Meaning $\mathcal{Y} = \{1, 2, \ldots, M\}$.

→ Solution: Re-define h^* as followed

$$p(x) = \begin{bmatrix} p_1(x) \\ \vdots \\ p_M(x) \end{bmatrix} \quad \text{where } p_c(x) = P(Y=c|X=x).$$

and let $h^*(x) = \underset{c}{\operatorname{argmax}} \{p_c(x)\}$.

$\Rightarrow$ We have that: $\sum_{y \in \mathcal{Y}} p_y(x) = 1, \ \forall x \in \mathcal{X}$.

(i) Given an arbitrary classifier $h: \mathcal{X} \to \mathcal{Y}$, we have:

$$R(h) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\mathbb{1}(h(X) \neq Y)\right]\right]$$

$$= \mathbb{E}_{x\sim X}\left[\sum_{y\in\mathcal{Y}} \mathbb{1}(h(x) \neq y)\, P(Y=y \mid X=x)\right] = \mathbb{E}_{x\sim X}\left[\sum_{y\in\mathcal{Y}} \mathbb{1}(h(x) \neq y)\, p_y(x)\right]$$

$\Rightarrow$ Let $\hat{y} = h(x)$ for a given $x \in \mathcal{X}$. We have:

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

$\Rightarrow$ To minimize $R(h)$, we need to maximize $p_{\hat{y}}(x)$ for all $x \in \mathcal{X}$. Hence, we have:

$$\hat{y} = h^*(x) = \operatorname{argmax}_c \{p_c(x)\} \Rightarrow h^* = \inf_h \{R(h)\}. \square$$

(iii) From (i) we have:

$$R^* = \mathbb{E}_X\left[1 - \max(\{p_c(x) : c \in \mathcal{Y}\})\right].$$

$\Rightarrow$ In other words, if we set $\overline{p_c(x)} = P(Y \neq c \mid X = x)$. We have:

$$R^* = \mathbb{E}_X\left[\min(\{\overline{p_c(x)} : c \in \mathcal{Y}\})\right]. \square$$ (This is similar to the formula for binary classification case).

Statistical Learning Theory – Exercise #1 (Chapters #1,2 → Continued[1]...)

① Problem 1 (Continued...)

(ii) Compute the excess risk $R(h) - R^*$:

$$R(h) - R^* = \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y)\, p_y(x)\right] - \mathbb{E}_{x \sim X}\left[1 - \max(\{p_c(x) : c \in \mathcal{Y}\})\right]$$

$$= \mathbb{E}_{x \sim X}\left[\sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y)\, p_y(x) + \max(\{p_c(x) : c \in \mathcal{Y}\}) - 1\right]$$

→ When $h(x) = h^*(x)$ $\Rightarrow$ $\sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y)\, p_y(x) + \max(\{p_c(x) : c \in \mathcal{Y}\}) = 1$

$$\Rightarrow \sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y)\, p_y(x) + \max(\{p_c(x) : c \in \mathcal{Y}\}) - 1 = 0$$

→ When $h(x) \neq h^*(x)$: Let $\hat{y} = h(x)$ and $\hat{y}_* = h^*(x)$. We have

$$\sum_{y \in \mathcal{Y}} \mathbb{1}(h(x) \neq y)\, p_y(x) + \max(\{p_c(x) : c \in \mathcal{Y}\}) - 1$$

$$= \sum_{y \neq \hat{y}} p_y(x) + p_{\hat{y}_*}(x) - 1$$

$$= 2p_{\hat{y}_*}(x) - 1 + \sum_{\substack{y \in \mathcal{Y};\, y \neq \hat{y};\\ y \neq \hat{y}_*}} p_y(x) = 2p_{\hat{y}_*}(x) - \Big(1 - \sum_{\substack{y \in \mathcal{Y};\, y \neq \hat{y};\\ y \neq \hat{y}_*}} p_y(x)\Big)$$

$$= 2p_{\hat{y}_*}(x) - P(Y \in \{\hat{y}, \hat{y}_*\} \mid X = x)$$

$$\Rightarrow R(h) - R^* = \mathbb{E}_{x \sim X}\Big[\Big(2p_{\hat{y}_*}(x) - \underbrace{P(Y \in \{\hat{y}, \hat{y}_*\} \mid X = x)}_{= p_{\hat{y}}(x) + p_{\hat{y}_*}(x).}\Big)\, \mathbb{1}(h(x) \neq h^*(x))\Big]$$

$$= \mathbb{E}_{x \sim X}\Big[\big(p_{\hat{y}_*}(x) - p_{\hat{y}}(x)\big)\, \mathbb{1}(h(x) \neq h^*(x))\Big]. \square$$

② **Problem 2.**

↳ Problem: Let $\alpha \in (0,1)$. Define the α-cost-sensitive risk of $h: \mathcal{X} \to \mathcal{Y}$ to be:

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

→ Determine the Bayes Classifier and prove an analogue of Theorem 1.1 for this risk.

→ Solution:

(i) Determine the Bayes Classifier & Bayes Risk:

→ We have: $R_\alpha(h) = \mathbb{E}_{x\sim X}\left[p(x)(1-\alpha)\mathbb{1}(y=1, h(x)=0) + \alpha(1-p(x))\mathbb{1}(y=0, h(x)=1)\right]$

⇒ We have the following table of values for the integrand inside $\mathbb{E}_{x\sim X}[\cdots]$:

$h(x)$ \ y	0	1
0	0	$p(x)(1-\alpha)$
1	$\alpha(1-p(x))$	0

⇒ For $R_\alpha(h)$ to be minimize, we have:

$$h^*(x) = \begin{cases} 1 & \text{when } \alpha(1-p(x)) \leq p(x)(1-\alpha) \\ 0 & \text{Otherwise.} \end{cases}$$

⇒ The Bayes Classifier is then defined as:

$$h^*(x) = \begin{cases} 1 & \text{when } p(x) \geq \alpha. \\ 0 & \text{Otherwise.} \end{cases}$$

⇒ We have the following Bayes Risk:

$$R^*_\alpha = R_\alpha(h^*) = \mathbb{E}_X\left[\min\left(\alpha(1-p(X)),\ p(X)(1-\alpha)\right)\right]. \square$$

② Problem 2 (Continued...)

(ii) Compute the Excess Risk: For all $h: \mathcal{X} \to \mathcal{Y}$, we have

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

$$= \mathbb{E}_{x\sim X}\left[p(x)(1-\alpha)\mathbb{1}(y=1, h(x)=0) + (1-p(x))\alpha\,\mathbb{1}(y=0, h(x)=1)\right]$$

$$= \mathbb{E}_{x\sim X}\left[p(x)(1-\alpha)\mathbb{1}(h(x)=0) + (1-p(x))\alpha\,\mathbb{1}(h(x)=1)\right]$$

(Because $\mathbb{1}(h(x)=0, y=0)$ and $\mathbb{1}(h(x)=1, y=1)$ carries no risk).

$$\Rightarrow R_\alpha(h) - R_\alpha^* = \mathbb{E}_{x\sim X}\Big[p(x)(1-\alpha)\left[\mathbb{1}(h(x)=0) - \mathbb{1}(h^*(x)=0)\right] + (1-p(x))\alpha\left[\mathbb{1}(h(x)=1) - \mathbb{1}(h^*(x)=1)\right]\Big]$$

$$= \mathbb{E}_{x\sim X}\Big[p(x)(1-\alpha)\left[\mathbb{1}(h(x)=0, h^*(x)=1) - \mathbb{1}(h(x)=1, h^*(x)=0)\right] + (1-p(x))\alpha\left[\mathbb{1}(h(x)=1, h^*(x)=0) - \mathbb{1}(h(x)=0, h^*(x)=1)\right]\Big]$$

$$= \mathbb{E}_{x\sim X}\Big[\mathbb{1}(h(x)=0, h^*(x)=1)\left[p(x)(1-\alpha) - \alpha(1-p(x))\right] + \mathbb{1}(h(x)=1, h^*(x)=0)\left[\alpha(1-p(x)) - p(x)(1-\alpha)\right]\Big]$$

$$= \mathbb{E}_{x\sim X}\Big[\mathbb{1}(h(x)=0, h^*(x)=1)(p(x)-\alpha) + \mathbb{1}(h(x)=1, h^*(x)=0)(\alpha - p(x))\Big]$$

$$= \mathbb{E}_X\left[|p(X)-\alpha|\,\mathbb{1}(h(X)\neq h^*(X))\right]. \quad \square$$

③ Problem 3.

↳ Corollary 1.2: $$R(\hat{h}_n) - R^* \leq 2\mathbb{E}_X\left[\left|p(X) - \hat{p}_n(X)\right|\right]$$

↳ Problem: Prove Corollary 1.2.

→ Solution: By Theorem 1.1, we have

$$R(\hat{h}_n) - R^* = 2\mathbb{E}_X\left[\left|p(X) - \tfrac{1}{2}\right| \mathbb{1}\left(\hat{h}_n(X) \neq h^*(X)\right)\right]$$

→ We have that $\hat{h}_n(X) \neq h^*(X)$ when:

(1) $\hat{p}_n(X) < \frac{1}{2}$ and $p(X) \geq \frac{1}{2}$

(2) or: $\hat{p}_n(X) \geq \frac{1}{2}$ and $p(X) < \frac{1}{2}$.

→ For (1) we have $p(X) - \frac{1}{2} \leq p(X) - \hat{p}_n(X)$. (Both sides positive)

→ For (2) we have $p(X) - \frac{1}{2} \geq p(X) - \hat{p}_n(X)$. (Both sides negative)

⇒ For both cases (1) and (2), we have $\left|p(X) - \hat{p}_n(X)\right| \geq \left|p(X) - \frac{1}{2}\right|$

⇒ We have: $2\mathbb{E}_X\left[\left|p(X) - \hat{p}_n(X)\right|\right]$

$$= 2\mathbb{E}_X\left[\left|p(X) - \hat{p}_n(X)\right|\left[\mathbb{1}\left(\hat{h}_n(X) \neq h^*(X)\right) + \mathbb{1}\left(\hat{h}_n(X) = h^*(X)\right)\right]\right]$$

$$\geq 2\mathbb{E}_X\left[\left|p(X) - \hat{p}_n(X)\right| \mathbb{1}\left(\hat{h}_n(X) \neq h^*(X)\right)\right]$$

$$\geq 2\mathbb{E}_X\left[\left|p(X) - \tfrac{1}{2}\right| \mathbb{1}\left(\hat{h}_n(X) \neq h^*(X)\right)\right] = R(\hat{h}_n) - R^*. \quad \square$$