

Exercise 2

Yawei Li

yawei.li@stat.uni-muenchen.de

Date

Exercise 1: Risk Minimizers for 0-1-Loss

Consider the classification learning setting, i.e., $\mathcal{Y} = \{1, \dots, g\}$, and the hypothesis space is $\mathcal{H} = \{h : \mathcal{X} \rightarrow \mathcal{Y}\}$. The loss function of interest is the 0-1-loss:

$$L(y, h(\mathbf{x})) = I_{y \neq h(\mathbf{x})} = \begin{cases} 1, & \text{if } y \neq h(\mathbf{x}), \\ 0, & \text{if } y = h(\mathbf{x}). \end{cases}$$

(a) Consider the hypothesis space of constant models

$\mathcal{H} = \{h : \mathcal{X} \rightarrow \mathcal{Y} | h(\mathbf{x}) = \theta \in \mathcal{Y} \forall \mathbf{x} \in \mathcal{X}\}$, where \mathcal{X} is the feature space. Show that

$$\hat{h}(\mathbf{x}) = \text{mode} \left\{ y^{(i)} \right\}$$

is the empirical risk minimizer for the 0-1-loss in this case.

Solution to Question (a)

The empirical risk is

$$\begin{aligned}\mathcal{R}_{\text{emp}}(h) &= \sum_{i=1}^n I_{y^{(i)} \neq h(\mathbf{x}^{(i)})} \\ &= \sum_{i=1}^n 1 - I_{y^{(i)} = h(\mathbf{x}^{(i)})} \quad \triangleright\end{aligned}$$

Therefore

$$\begin{aligned}\arg \min_{h \in \mathcal{H}} \mathcal{R}_{\text{emp}}(h) &= \arg \min_{h \in \mathcal{H}} \sum_{i=1}^n 1 - I_{y^{(i)} = h(\mathbf{x}^{(i)})} \\ &= \arg \max_{h \in \mathcal{H}} \sum_{i=1}^n I_{y^{(i)} = h(\mathbf{x}^{(i)})} \\ &= \arg \max_{\theta \in \mathcal{Y}} \sum_{i=1}^n I_{y^{(i)} = \theta} = \text{mode} \{y^{(i)}\}\end{aligned}$$

Question (b)

(b) What is the optimal constant model in terms of the (theoretical) risk for the 0-1-loss and what is its risk?

Solution to Question (b)

$$\begin{aligned}\mathcal{R}_L(h) &= \int_y \int_{\mathbf{x}} I_{y \neq h(\mathbf{x})} p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \int_y \int_{\mathbf{x}} I_{y \neq \theta} p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \int_y I_{y \neq \theta} p(y) dy \\ &= \int_y (1 - I_{y=\theta}) p(y) dy\end{aligned}$$

Therefore, $\arg \min_h \mathcal{R}_L(h) = \arg \max_{\theta \in \mathcal{Y}} \int_y I_{y=\theta} p(y) dy$. Furthermore, $\mathcal{Y} = \{1, \dots, g\}$, it follows that $\arg \max_{\theta \in \mathcal{Y}} \int_y I_{y=\theta} p(y) dy = \arg \max_{\theta \in \mathcal{Y}} \sum_{j=1}^g I_{\theta=j} \mathbb{P}(y = j)$. (Show example.)

Hence, the optimal constant model for the **theoretical** risk is

$$\bar{h}(\mathbf{x}) = \arg \max_{l \in \mathcal{Y}} \mathbb{P}(y = l)$$

Solution to Question (b): Continued

Before we compute $\mathcal{R}_L(\bar{h})$, we write 0-1-loss as follows:

$$L(y, h(\mathbf{x})) = I_{y \neq h(\mathbf{x})} = \sum_{k \in \mathcal{Y}} I_{y=k} I_{k \neq h(\mathbf{x})} = \sum_{k \in \mathcal{Y}} L(k, h(\mathbf{x})).$$

Then, the risk of \bar{h} is

$$\begin{aligned} \mathcal{R}_L(\bar{h}) &= \mathbb{E}_{xy} [L(y, \bar{h}(\mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}} [\mathbb{E}_{y|\mathbf{x}} [L(y, \bar{h}(\mathbf{x})) \mid \mathbf{x}]] \\ &= \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{y|\mathbf{x}} \left[\sum_{k \in \mathcal{Y}} I_{y=k} L(k, \bar{h}(\mathbf{x})) \mid \mathbf{x} \right] \right] \\ &= \mathbb{E}_{\mathbf{x}} \left[\sum_{k \in \mathcal{Y}} L(k, \bar{h}(\mathbf{x})) \mathbb{E}_{y|\mathbf{x}} [I_{y=k} \mid \mathbf{x}] \right] \quad \triangleright \end{aligned}$$

Solution to Question (b): Continued

$$\begin{aligned}\mathcal{R}_L(\bar{h}) &= \mathbb{E}_{\mathbf{x}} \left[\sum_{k \in \mathcal{Y}} L(k, \bar{h}(\mathbf{x})) \mathbb{E}_{y|\mathbf{x}} [I_{y=k} \mid \mathbf{x}] \right] \\&= \mathbb{E}_{\mathbf{x}} \left[\sum_{k \in \mathcal{Y}} L(k, \bar{h}(\mathbf{x})) \mathbb{P}(y = k \mid \mathbf{x}) \right] \\&= \sum_{k \in \mathcal{Y}} L(k, \bar{h}(\mathbf{x})) \mathbb{E}_{\mathbf{x}} [\mathbb{P}(y = k \mid \mathbf{x})] \\&= \sum_{k \in \mathcal{Y}} L(k, \bar{h}(\mathbf{x})) \mathbb{P}(y = k) \\&= \sum_{k \in \mathcal{Y}} I_{k \neq \bar{h}(\mathbf{x})} \mathbb{P}(y = k) \\&= \sum_{k \in \mathcal{Y}} I_{k \neq \arg \max_{l \in \mathcal{Y}} \mathbb{P}(y=l)} \mathbb{P}(y = k) \\&= 1 - \max_{l \in \mathcal{Y}} \mathbb{P}(y = l).\end{aligned}$$

Question (c)

(c) Derive the approximation error if the hypothesis space \mathcal{H} consists of the **constant models**.

Recall that the approximation error is defined as

$$\inf_{h \in \mathcal{H}} \mathcal{R}_L(h) - \mathcal{R}_L^*$$

Solution to (c)

$$\begin{aligned}\inf_{h \in \mathcal{H}} \mathcal{R}_L(h) - \mathcal{R}_L^* &= \mathcal{R}_L(\bar{h}) - \mathcal{R}_L^* \\ &= (1 - \max_{l \in \mathcal{Y}} \mathbb{P}(y = l)) - (1 - \mathbb{E}_{\mathbf{x}}[\max_{l \in \mathcal{Y}} \mathbb{P}(y = l | \mathbf{x})]) \\ &= \mathbb{E}_{\mathbf{x}}[\max_{l \in \mathcal{Y}} \mathbb{P}(y = l | \mathbf{x})] - \max_{l \in \mathcal{Y}} \mathbb{P}(y = l).\end{aligned}$$

Question (d)

(d) Assume now $g = 2$ (binary classification) and consider now the hypothesis space of probabilistic classifiers $\mathcal{H} = \{\pi : \mathcal{X} \rightarrow [0, 1]\}$, that is, $\pi(\mathbf{x})$ (or $1 - \pi(\mathbf{x})$) is an estimate of the posterior distribution $p_{y|\mathbf{x}}(1|\mathbf{x})$ (or $p_{y|\mathbf{x}}(0|\mathbf{x})$). Furthermore, consider the probabilistic 0-1-loss

$$L(y, \pi(\mathbf{x})) = \begin{cases} 1, & \text{if } (\pi(\mathbf{x}) \geq 1/2 \text{ and } y = 0) \text{ or } (\pi(\mathbf{x}) < 1/2 \text{ and } y = 1), \\ 0, & \text{else.} \end{cases}$$

Is the minimum of $\mathbb{E}_{xy}[L(y, \pi(\mathbf{x}))]$ unique over $\pi \in \mathcal{H}$? Is the posterior distribution $p_{y|x}$ a resp. the minimizer of $\mathbb{E}_{xy}[L(y, \pi(\mathbf{x}))]$? Discuss the corresponding (dis-)advantages of your findings.

Solution to Question (d)

- We can rewrite the 0-1-loss as

$$L(y, \pi(\mathbf{x})) = I_{\pi(\mathbf{x}) \geq 1/2} I_{y=0} + I_{\pi(\mathbf{x}) < 1/2} I_{y=1}.$$

- Since $\mathcal{H} = \{\pi : \mathcal{X} \rightarrow [0, 1]\}$, we can optimize π for each point \mathbf{x} .
- In other words, for $\mathbb{E}_{xy}[L(y, \pi(\mathbf{x}))] = \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[L(y, \pi(\mathbf{x})) | \mathbf{x}]]$, we optimize $\mathbb{E}_{y|\mathbf{x}}[L(y, \pi(\mathbf{x})) | \mathbf{x}]$ for each \mathbf{x} .

Solution to Question (d): Continued

$$\begin{aligned}\mathbb{E}_{y|\mathbf{x}}[L(y, \pi(\mathbf{x})) \mid \mathbf{x}] &= \mathbb{E}_{y|\mathbf{x}}[I_{\pi(\mathbf{x}) \geq 1/2} I_{y=0} + I_{\pi(\mathbf{x}) < 1/2} I_{y=1} \mid \mathbf{x}] \\&= \mathbb{E}_{y|\mathbf{x}}[I_{\pi(\mathbf{x}) \geq 1/2} I_{y=0} \mid \mathbf{x}] + \mathbb{E}_{y|\mathbf{x}}[I_{\pi(\mathbf{x}) < 1/2} I_{y=1} \mid \mathbf{x}] \\&= I_{\pi(\mathbf{x}) \geq 1/2} \cdot \mathbb{E}_{y|\mathbf{x}}[I_{y=0} \mid \mathbf{x}] + I_{\pi(\mathbf{x}) < 1/2} \cdot \mathbb{E}_{y|\mathbf{x}}[I_{y=1} \mid \mathbf{x}] \quad \triangleright \\&= I_{\pi(\mathbf{x}) \geq 1/2} \mathbb{P}(y = 0 \mid \mathbf{x}) + I_{\pi(\mathbf{x}) < 1/2} \mathbb{P}(y = 1 \mid \mathbf{x}).\end{aligned}$$

Solution to Question (d): Continued

$$\mathbb{E}_{y|\mathbf{x}}[L(y, \pi(\mathbf{x})) | \mathbf{x}] = I_{\pi(\mathbf{x}) \geq 1/2} \mathbb{P}(y = 0 | \mathbf{x}) + I_{\pi(\mathbf{x}) < 1/2} \mathbb{P}(y = 1 | \mathbf{x}).$$

We can distinguish between two cases:

- ▶ If $\mathbb{P}(y = 0 | \mathbf{x}) \geq 1/2$, then any $\pi(\mathbf{x}) < 1/2$ minimizes $\mathbb{E}_{y|\mathbf{x}}[L(y, \pi(\mathbf{x})) | \mathbf{x}]$.
- ▶ If $\mathbb{P}(y = 0 | \mathbf{x}) \leq 1/2$, then any $\pi(\mathbf{x}) \geq 1/2$ minimizes $\mathbb{E}_{y|\mathbf{x}}[L(y, \pi(\mathbf{x})) | \mathbf{x}]$.

In other words:

$$\pi(\mathbf{x}) = \begin{cases} < 1/2, & \text{if } \mathbb{P}(y = 0 | \mathbf{x}) \geq 1/2, \\ \geq 1/2, & \text{if } \mathbb{P}(y = 1 | \mathbf{x}) < 1/2. \end{cases}$$

The posterior distribution $p_{y|\mathbf{x}}(1 | \mathbf{x})$ is quite naturally of this form, but it is not the only π of this kind. As a consequence, the minimize is not unique.

Solution to Question (d): Continued

Disadvantages of using $p_{y|x}$:

- ▶ TODO (The solution is not very clear. Ask people around.)