Homework 1

Shao-Ting Chiu (UIN:433002162)

9/19/22

Homework Description

Course: ECEN649, Fall2022

Problems (from Chapter 2 in the book): 2.1 , 2.3 (a,b), 2.4, 2.7, 2.9, 2.17 (a,b)

Note: the book is available electronically on the Evans library website.

- Deadline: `Sept. 26th, 11:59 pm`

Problem 2.1

Suppose that X is a discrete feature vector, with distribution concentrated over a countable set $D = \{x^1, x^2, \dots\}$ in R^d. Derive the discrete versions of (2.3), (2.4), (2.8), (2.9), (2.11), (2.30), (2.34), and (2.36)

Hint: Note that if X has a discrete distribution, then integration becomes summation, $P(X = x_k)$, for $x_k \in D$, play the role of $p(x)$, and $P(X = x_k|Y = y)$, for $x_k \in D$, play the role of $p(x|Y = y)$, for $y = 0, 1$.

(2.3)

From Braga-Neto (2020, 16)

$$P(X \in E, Y = 0) = \int_E P(Y = 0)p(x|Y = 0)dx \tag{1}$$

$$P(X \in E, Y = 1) = \int_E P(Y = 1)p(x|Y = 1)dx \tag{2}$$

$$\tag{3}$$

Let $x_k = [x_1, \dots, x_d]$ be the feature vector of X in set $D \in R^d$,

$$\begin{aligned}
P(X \in D, Y = 0) &= P(X = [x_1, \dots, x_d], Y = 0) && (4)\\
&= \sum_{X \in D} P(Y = 0) P(X = [x_1, \dots, x_d] | Y = 0) && (5)\\
P(X \in D, Y = 1) &= P(X = [x_1, \dots, x_d], Y = 1) && (6)\\
&= \sum_{X \in D} P(Y = 1) P(X = [x_1, \dots, x_d] | Y = 1) && (7)\\
& && (8)
\end{aligned}$$

(2.4)

From Braga-Neto (2020, 17)

$$\begin{aligned}
P(Y = 0 | X = x_k) &= \frac{P(Y = 0) p(X = x_k | Y = 0)}{p(X = x_k)} && (9)\\
&= \frac{P(Y = 0) p(X = x_k | Y = 0)}{P(Y = 0) p(X = x_k | Y = 0) + P(Y = 1) p(X = x_k | Y = 1)} && (10)\\
& && (11)
\end{aligned}$$

$$\begin{aligned}
P(Y = 1 | X = x_k) &= \frac{P(Y = 1) p(X = x_k | Y = 1)}{p(X = x_k)} && (12)\\
&= \frac{P(Y = 1) p(X = x_k | Y = 1)}{P(Y = 0) p(X = x_k | Y = 0) + P(Y = 1) p(X = x_k | Y = 1)} && (13)\\
& && (14)
\end{aligned}$$

(2.8)

From Braga-Neto (2020, 18)

$$\epsilon^0[\psi] = P(\psi(X) = 1 | Y = 0) = \sum_{\{x_k | \psi(x) = 1\}} p(x_k | Y = 0)$$

$$\epsilon^1[\psi] = P(\psi(X) = 0 | Y = 1) = \sum_{\{x_k | \psi(x) = 1\}} p(x_k | Y = 1)$$

(2.9)

From Braga-Neto (2020, 18)

$$\epsilon[\psi] = \sum_{\{x|\psi(x)=1\}} P(Y=0)p(x_k|Y=0) + \sum_{\{x|\psi=0\}} P(Y=1)p(x_k|Y=1)$$

(2.11)

From Braga-Neto (2020, 19)

$$\epsilon[\psi] = E[\epsilon[\psi|X=x_k]] = \sum_{x_k \in D} \epsilon[\psi|X=x_k]p(x_k)$$

(2.30)

(2.34)

(2.36)

Problem 2.3

This problem seeks to characterize the case $\epsilon^* = 0$.

(a)

Prove the "Zero-One Law" for perfect discrimination:

$$\epsilon^* = 0 \Leftrightarrow \eta(X) = 0 \text{ or } 1 \quad \text{with probability 1.} \tag{15}$$

The optimal Bayes classifier is defined in Braga-Neto (2020, 20). That is

$$\psi^*(x) = \arg\max_i P(Y=i|X=x) = \begin{cases} 1, & \eta(x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Part 1: $\eta(X) = 1$

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

$$\because \eta(X) = 1 > \frac{1}{2} \therefore \psi^*(x) = 1$$

$$\begin{aligned}
\epsilon^* &= \epsilon[\psi^*(X)|X=x] && (17)\\
&= I_{\psi^*(x)=0} P(Y=1|X=x) + I_{\psi^*(x)=1} P(Y=0|X=x) && (18)\\
&= \underbrace{I_{\psi^*(x)=0}}_{=0} \underbrace{\eta(X)}_{=1} + \underbrace{I_{\psi^*(x)=1}}_{=1} \underbrace{(1-\eta(X))}_{=0} && (19)\\
&= 0 && (20)
\end{aligned}$$

Part 2: $\eta(X) = 0$

Similarly,

$$\because \eta(X) = 0 \leq \frac{1}{2} \therefore \psi^*(x) = 0$$

$$\begin{aligned}
\epsilon^* &= \epsilon[\psi^*(X)|X=x] && (21)\\
&= I_{\psi^*(x)=0} P(Y=1|X=x) + I_{\psi^*(x)=1} P(Y=0|X=x) && (22)\\
&= \underbrace{I_{\psi^*(x)=0}}_{=1} \underbrace{\eta(X)}_{=0} + \underbrace{I_{\psi^*(x)=1}}_{=0} \underbrace{(1-\eta(X))}_{=1} && (23)\\
&= 0 && (24)
\end{aligned}$$

In conclusion, both cases shows that $\epsilon^* = 0$.

(b)

Show that

$$\epsilon^* = 0 \Leftrightarrow \text{ there is a function } f \text{ s.t. } Y = f(X) \text{ with probability 1}$$

$$\eta(X) = Pr(Y=1|X=x) = \begin{cases} 1, & f(X) = 1 \\ 0, & f(X) = 0 \end{cases} \tag{25}$$

The sceneraio is same as Problem 3.7 (a).

1. Given $\eta(X) = 1$

- $\epsilon^* = 0$

2. Given $\eta(X) = 0$

   - $\epsilon^* = 0$

$\epsilon^* = 0$ for both cases.

Problem 2.4

This problem concerns the extension to the multiple-class case of some of the concepts derived in this chapter. Let $Y \in \{0, 1, \ldots, c-1\}$, where c is the number of classes, and let

$$\eta_i(x) = P(Y = i|X = x), \quad i = 0, 1, \ldots, c-1,$$

for each $x \in R^d$. We need to remember that these probabilities are not indpendent, but satisfy $\eta_0(x) + \eta_1(x) + \cdots + \eta_{c-1}(x) = 1$, for each $x \in R^d$, so that one of the functions is redundant. In the two-class case, this is made explicit by using a single $\eta(x)$, but using the redundant set above proves advantageous in the multiple-class case, as seen below.

Hint: you should answer the following items in sequence, using the previous answers in the solution of the following ones

(a)

Given a classifier $\psi : R^d \to \{0, 1, \ldots, c-1\}$, show that its conditional error $P(\psi(X) \neq Y|X = x)$ is given by

$$P(\psi(X) \neq Y|X = x) = 1 - \sum_{i=1}^{c-1} I_{\psi(x)=i}\eta_i(x) = 1 - \eta_{\psi(x)}(x) \tag{26}$$

Use the "Law of Total Probability" (Braga-Neto 2020, sec. A.53),

$$P(\psi(X) = Y|X = x) + P(\psi(X) \neq Y|X = x) = 1 \tag{27}$$

$\therefore$ We can derive the probability of error via

$$
\begin{aligned}
P(\psi(X) \neq Y|X = x) &= 1 - P(\psi(X) = Y|X = x) &(28)\\
&= 1 - \sum_{i=0}^{c-1} P(\psi(x) = i, Y = i|X = x) &(29)\\
&= 1 - \sum_{i=0}^{c-1} I_{\psi(x)=i} P(Y = i|X = x) &(30)\\
&= 1 - \sum_{i=0}^{c-1} I_{\psi(x)=i} \eta_i(x) &(31)
\end{aligned}
$$

Combining together, Equation 27 implies Equation 26.

(b)

Assuming that X has a density, show that the classification error of ψ is given by

$$\epsilon = 1 - \sum_{i=0}^{c-1} \int_{\{x|\psi(x)=i\}} \eta_i(x)p(x)dx.$$

Let $\{x|\psi(x) = i\}$ be the set of $\psi(x) = i$ in X.

Use the *multiplication rule* (Braga-Neto 2020, sec. A1.3)

$$
\begin{aligned}
\epsilon &= E[\epsilon[\psi(x)|X = x]] &(32)\\
&= 1 - \int_{R^d} P(\psi(X) = Y|X = x)p(x)dx &(33)\\
&= 1 - \sum_{i=0}^{c-1} \int_{R^d} p(\psi(X) = i, Y = i|X = x)p(x)dx &(34)\\
&= 1 - \sum_{i=0}^{c-1} \int_{R^d} \underbrace{p(\psi(X) = i|X = x)}_{=1 \text{ if } \{x|\psi(x)=i\}; 0, \text{ otherwise.}} p(Y = i|X = x)p(x)dx &(35)\\
&= 1 - \sum_{i=0}^{c-1} \int_{\{x|\psi(x)=i\}} 1 \cdot p(Y = i|X = x)p(x)dx &(36)\\
&= 1 - \sum_{i=0}^{c-1} \int_{\{x|\psi(x)=i\}} p(Y = i|X = x)p(x)dx &(37)
\end{aligned}
$$

(c)

Prove that the Bayes classifier is given by

$$\psi^*(x) = \arg \max_{i=0,1,\ldots,c-1} \eta_i(x), \quad x \in R^d$$

Hint: Start by considering the difference between conditional expected errors $P(\psi(X) \neq Y|X = x) - P(\psi^*(X) \neq Y|X = x)$.

(d)

Show that the Bayes error is given by

$$\epsilon^* = 1 - E[\max_{i=0,1,\ldots,c-1} \eta_i(X)]$$

(e)

Show that the maximum Bayes error possible is $1 - \frac{1}{c}$.

Problem 2.7

Consider the following univariate Gaussian class-conditional densities:

$$p(x|Y = 0) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-3)^2}{2})$$

$$p(x|Y = 1) = \frac{1}{3\sqrt{2\pi}} \exp(-\frac{(x-4)^2}{18})$$

Assume that the classes are equally likely, i.e., $P(Y = 0) = P(Y = 1) = \frac{1}{2}$

(a)

Draw the densities and determine the Bayes classifier graphically.

(b)

Determine the Bayes classifier.

(c)

Determine the specificity and sensitivity of the Bayes classifier.

Hint: use the standard Gaussian CDF $\psi(x)$

Table 1: The definition of sensitivity and specificity from Braga-Neto (2020, 18)

Sensitivity	Specificity
$1 - \epsilon^1[\psi]$	$1 - \epsilon^0[\psi]$

(d)

Determine the overall Bayes error.

Problem 2.9

Obtain the optimal decision boundary in the Gaussian model with $P(Y = 0) = P(Y = 1)$ and

In each case draw the optimal decision boundary, along with the class means and class conditional density contours, indicating the 0- and 1-decision regions.

(a)

$$\mu_0 = (0, 0)^T, \mu_1 = (2, 0)^T, \Sigma_0 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$$

(b)

$$\mu_0 = (0, 0)^T, \mu_1 = (2, 0)^T, \Sigma_0 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

(c)

$$\mu_0 = (0, 0)^T, \mu_1 = (0, 0)^T, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

(d)

$$\mu_0 = (0,0)^T, \mu_1 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

Python Assignment: Problem 2.17

This problem concerns the Gaussian model for synthetic data generation in Braga-Neto (2020, sec. A8.1).

(a)

Derive a general expression for the Bayes error for the homoskedastic case with $\mu_0 = (0, \dots, 0), \mu_1 = (1, \dots, 1)$, and $P(Y = 0) = P(Y = 1)$. Your answer should be in terms of $k, \sigma_1^2, \dots, \sigma_k^2, l_1, \dots, l_k$, and $\sigma_1, \dots, \sigma_k$.

Hint: Use the fact that

$$\begin{bmatrix} 1 & \sigma & \cdots & \sigma \\ \sigma & 1 & \cdots & \sigma \\ \vdots & \vdots & \ddots & \vdots \\ \sigma & \sigma & \cdots & 1 \end{bmatrix}_{l\times l}^{-1} = \frac{1}{(1-\sigma)(1+(l-1)\sigma)} \begin{bmatrix} 1+(l-2)\sigma & -\sigma \cdots -\sigma & & \\ -\sigma & 1+(l-2)\sigma & \cdots -\sigma & \\ \vdots & \vdots & \ddots & \vdots \\ -\sigma & -\sigma & \cdots & 1+(l-2)\sigma \end{bmatrix} \tag{38}$$

(b)

Specialize the previous formula for equal-sized blocks $l_1 = \cdots = l_k = l$ with equal correlations $\sigma_1 = \cdots = \sigma_k = \sigma$, and constant variance $\sigma_1^2 = \cdots, \sigma_k^2 = \sigma^2$. Write the resulting formula in terms of d, l, σ and σ.

i.

Using the python function `norm.cdf` in the `scipy.stats` module, plot the Bayes error as a function of $\sigma \in [0.01, 3]$ for $d = 20, l = 4$, and four different correlation values $\sigma = 0, 0.25, 0.5, 0.75$ (plot one curve for each value). Confirm that the Bayes error increasese monotonically with σ from 0 to 0.5 for each value of σ, and that Bayes error for large σ is uniformly larger than that for smaller σ. The latter fact shows that correlation between the features is detrimental to classfification.

ii.

Plot the Bayes error as a function of $d = 2, 4, 6, 8, \ldots, 40$, with fixed block size $l = 4$ and variance $\sigma^2 = 1$ and $\sigma = 0, 0.25, 0.5, 0.75$ (plot one curve for each value). Confirm that the Bayes error decreases monotonically to 0 with increasing dimensionality, with faster convergence for smaller correlation values.

iii.

Plot the Bayes error as a function of the correlation $\sigma \in [0, 1]$ for constant variance $\sigma^2 = 2$ and fixed $d = 20$ with varying block size $l = 1, 2, 4, 10$ (plot one curve for each value). Confirm that the Bayes error increases monotonically with increasing correlation. Notice that the rate of increase is particularly large near $\sigma = 0$, which shows that the Bayes error is very sensitive to correlation in the near-independent region.

References

Braga-Neto, Ulisses. 2020. *Fundamentals of Pattern Recognition and Machine Learning.* Springer.