

Homework 1

Shao-Ting Chiu (UIN:433002162)

9/23/22

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

Computational Environment Setup

Third-party libraries

```
1  %matplotlib inline
2  import sys # system information
3  import matplotlib # plotting
4  import scipy as st # scientific computing
5  import pandas as pd # data managing
6  import numpy as np # numerical computation
7  import scipy.optimize as opt
8  import matplotlib.pyplot as plt
9  from scipy.special import erf
10 from numpy.linalg import inv, det
11 from scipy.linalg import block_diag
12 from scipy.stats import norm # for problem 2.17 (b)
13 from scipy.stats import multivariate_normal
14 # Matplotlib setting
15 plt.rcParams['text.usetex'] = True
16 matplotlib.rcParams['figure.dpi']= 300
```

```
17 RES_GRID = 150
```

Version

```
1 print(sys.version)
2 print(matplotlib.__version__)
3 print(st.__version__)
4 print(np.__version__)
5 print(pd.__version__)
```

3.8.12 (default, Oct 22 2021, 18:39:35)

[Clang 13.0.0 (clang-1300.0.29.3)]

3.3.1

1.5.2

1.19.1

1.1.1

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)

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

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

$$(3)$$

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

$$P(X \in D, Y = 0) = P(X = [x_1, \dots, x_d], Y = 0) \quad (4)$$

$$= \sum_{X_k \in D} P(Y = 0)P(X = [x_1, \dots, x_d]|Y = 0) \quad (5)$$

$$P(X \in D, Y = 1) = P(X = [x_1, \dots, x_d], Y = 1) \quad (6)$$

$$= \sum_{X_k \in D} P(Y = 1)P(X = [x_1, \dots, x_d]|Y = 1) \quad (7)$$

$$(8)$$

(2.4)

From Braga-Neto (2020, 17)

$$P(Y = 0|X = x_k) = \frac{P(Y = 0)p(X = x_k|Y = 0)}{p(X = x_k)} \quad (9)$$

$$= \frac{P(Y = 0)p(X = x_k|Y = 0)}{P(Y = 0)p(X = x_k|Y = 0) + P(Y = 1)p(X = x_k|Y = 1)} \quad (10)$$

$$(11)$$

$$P(Y = 1|X = x_k) = \frac{P(Y = 1)p(X = x_k|Y = 1)}{p(X = x_k)} \quad (12)$$

$$= \frac{P(Y = 1)p(X = x_k|Y = 1)}{P(Y = 0)p(X = x_k|Y = 0) + P(Y = 1)p(X = x_k|Y = 1)} \quad (13)$$

$$(14)$$

(2.8)

From Braga-Neto (2020, 18)

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

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

(2.9)

From Braga-Neto (2020, 18)

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

(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)

From Braga-Neto (2020, 24).

$$\epsilon^* = \sum_{x_k \in X} \left[I_{\eta(X=x_k) \leq 1-\eta(X=x_k)} \eta(X=x_k) + I_{\eta(X=x_k) > 1-\eta(X=x_k)(1-\eta(X=x_k))} \right] p(X=x_k)$$

(2.34)

From Braga-Neto (2020, 25).

$$\epsilon^* = P(Y=0)\epsilon^0[\psi^*] + P(Y=1)\epsilon^1[\psi^*] \tag{15}$$

$$= \sum_{\{x_k | P(Y=1)p(x_k|Y=1) > P(Y=0)p(x_k|Y=0)\}} P(Y=0)p(x_k|Y=0) \tag{16}$$

$$+ \sum_{\{x_k | P(Y=1)p(x_k|Y=1) \leq P(Y=0)p(x_k|Y=0)\}} P(Y=1)p(x_k|Y=1) \tag{17}$$

(2.36)

From Braga-Neto (2020, 25)

$$E[\eta(X)] = \sum_{x_k \in R^d} P(Y = 1|X = x_k)p(x_k) = P(Y = 1)$$

Problem 2.3

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

(a)

Prove the “Zero-One Law” for perfect discrimination:

$$\epsilon^* = 0 \Leftrightarrow \eta(X) = 0 \text{ or } 1 \quad \text{with probability 1.} \quad (18)$$

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

$$\psi^*(x) = \arg \max_i P(Y = i|X = x) = \begin{cases} 1, & \eta(x) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \quad (19)$$

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

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

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

$$\epsilon^* = \epsilon[\psi^*(X)|X = x] \quad (20)$$

$$= I_{\psi^*(x)=0}P(Y = 1|X = x) + I_{\psi^*(x)=1}P(Y = 0|X = x) \quad (21)$$

$$= \underbrace{I_{\psi^*(x)=0}}_{=0} \underbrace{\eta(X)}_{=1} + \underbrace{I_{\psi^*(x)=1}}_{=1} \underbrace{(1 - \eta(X))}_{=0} \quad (22)$$

$$= 0 \quad (23)$$

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

Similarly,

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

$$\epsilon^* = \epsilon[\psi^*(X)|X = x] \tag{24}$$

$$= I_{\psi^*(x)=0}P(Y = 1|X = x) + I_{\psi^*(x)=1}P(Y = 0|X = x) \tag{25}$$

$$= \underbrace{I_{\psi^*(x)=0}}_{=1} \underbrace{\eta(X)}_{=0} + \underbrace{I_{\psi^*(x)=1}}_{=0} \underbrace{(1 - \eta(X))}_{=1} \tag{26}$$

$$= 0 \tag{27}$$

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

(b)

Show that

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

$$\eta(X) = Pr(Y = 1|X = x) = \begin{cases} 1, & f(X) = 1 \\ 0, & f(X) = 0 \end{cases} \tag{28}$$

The sceneraio is same as [Problem 3.7 \(a\)](#).

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

$$\bullet \epsilon^* = 0$$

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

$$\bullet \epsilon^* = 0$$

$\epsilon^* = 0$ 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 $Y \in \{0, 1, \dots, c-1\}$, where c is the number of classes, and let

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

for each $x \in R^d$. We need to remember that these probabilities are not independent, but satisfy $\eta_0(x) + \eta_1(x) + \dots + \eta_{c-1}(x) = 1$, 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 $\psi : R^d \rightarrow \{0, 1, \dots, c-1\}$, show that its conditional error $P(\psi(X) \neq Y|X = x)$ is given by

$$P(\psi(X) \neq Y|X = x) = 1 - \sum_{i=1}^{c-1} I_{\psi(x)=i} \eta_i(x) = 1 - \eta_{\psi(x)}(x) \quad (29)$$

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

$$P(\psi(X) = Y|X = x) + P(\psi(X) \neq Y|X = x) = 1 \quad (30)$$

\therefore We can derive the probability of error via

$$P(\psi(X) \neq Y|X = x) = 1 - P(\psi(X) = Y|X = x) \quad (31)$$

$$= 1 - \sum_{i=0}^{c-1} P(\psi(x) = i, Y = i|X = x) \quad (32)$$

$$= 1 - \sum_{i=0}^{c-1} I_{\psi(x)=i} P(Y = i|X = x) \quad (33)$$

$$= 1 - \sum_{i=0}^{c-1} I_{\psi(x)=i} \eta_i(x) \quad (34)$$

Combining together, Equation 30 implies Equation 29.

(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)

$$\epsilon = E[\epsilon[\psi(x)|X=x]] \tag{35}$$

$$= 1 - \int_{R^d} P(\psi(X) = Y | X = x) p(x) dx \tag{36}$$

$$= 1 - \sum_{i=0}^{c-1} \int_{R^d} p(\psi(X) = i, Y = i | X = x) p(x) dx \tag{37}$$

$$= 1 - \sum_{i=0}^{c-1} \int_{R^d} \underbrace{p(\psi(X) = i | X = x)}_{=1 \text{ if } \{x|\psi(x)=i\}; 0, \text{ otherwise.}} p(Y = i | X = x) p(x) dx \tag{38}$$

$$= 1 - \sum_{i=0}^{c-1} \int_{\{x|\psi(x)=i\}} 1 \cdot p(Y = i | X = x) p(x) dx \tag{39}$$

$$= 1 - \sum_{i=0}^{c-1} \int_{\{x|\psi(x)=i\}} p(Y = i | X = x) p(x) dx \tag{40}$$

(c)

Prove that the Bayes classifier is given by

$$\psi^*(x) = \arg \max_{i=0,1,\dots,c-1} \eta_i(x), \quad x \in R^d \tag{41}$$

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

According to Braga-Neto (2020, 20), a Bayes classifier (ψ^*) is defined as

$$\psi^* = \arg \min_{\psi \in \mathcal{C}} P(\psi(X) \neq Y)$$

over the set \mathcal{C} of all classifiers. We need to show that the error of any $\psi \in \mathcal{C}$ has the conditional error rate:

$$\epsilon[\psi|X = x] \geq \epsilon[\psi^*|X = x], \quad \text{for all } x \in R^d \quad (42)$$

From Equation 29, classifiers have the error rates:

$$P(\psi^*(X) \neq |X = x) = 1 - \sum_{i=1}^{c-1} I_{\psi^*(x)=i} \eta_i(x) \quad (43)$$

$$P(\psi(X) \neq |X = x) = 1 - \sum_{i=1}^{c-1} I_{\psi(x)=i} \eta_i(x) \quad (44)$$

Therefore,

$$P(\psi(X) \neq Y|X = x) - P(\psi^*(X) \neq Y|X = x) = (1 - \sum_{i=1}^{c-1} I_{\psi(x)=i} \eta_i(x)) - (1 - \sum_{i=1}^{c-1} I_{\psi^*(x)=i} \eta_i(x)) \quad (45)$$

$$= \sum_{i=1}^{c-1} (I_{\psi^*(x)=i} - I_{\psi(x)=i}) \eta_i(x) \quad (46)$$

\therefore

- $I_{\psi^*(x)=i^*} = 1$ when i^* satisfies $\eta_{i^*}(x) = \max_{i=0,1,\dots,c-1} \eta_i(x) = \eta_{\max}(x)$
- $I_{\psi(x)=i'} = 1$ when $\psi(x) = i'$ for $i' \in 0, 1, \dots, c-1$

\therefore

if $i^* \neq i'$

$$P(\psi(X) \neq Y|X = x) - P(\psi^*(X) \neq Y|X = x) = (1 - 0)\eta_{i^*}(x) + (0 - 1)\eta_{i'}(x) \quad (47)$$

$$= \eta_{i^*}(x) - \eta_{i'}(x) \quad (48)$$

$$= \eta_{\max}(x) - \eta_{i'}(x) \quad (49)$$

$$\geq 0 \quad (50)$$

if $i^* = i'$

$$P(\psi(X) \neq Y|X = x) - P(\psi^*(X) \neq Y|X = x) = \eta_{i^*}(x) - \eta_{i'}(x) = 0$$

Therefore, there is no classifier $\psi \in \mathcal{C}$ can have conditional error rate lower than Bayes classifier Equation 41.

(d)

Show that the Bayes error is given by

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

From Problem 2.4.b,

- Noted that, $\{x|\psi^*(x) = i\} = \emptyset$ if $i \neq i^*$

$$\epsilon[\psi^*] = E[\epsilon[\psi^*(x)|X = x]] \tag{51}$$

$$= 1 - \sum_{i=0}^{c-1} \int_{\{x|\psi^*(x)=i\}} \eta_i(x)p(x)dx \tag{52}$$

$$= 1 - \int_{\{x|\psi^*(x)=i^*\}} \eta_{\max}(x)p(x)dx \tag{53}$$

$$= 1 - E[\eta_{\max}(x)] \tag{54}$$

(e)

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

$$\max \epsilon[\psi^*] = 1 - \min E[\max_{i=0,1,\dots,c-1} \eta_i(X)] \tag{55}$$

also,

given

$$\eta_1(x) = \eta_2(x) = \dots = \eta_{c-1}(x)$$

Table 1: Parameters of Gaussian PDFs.

	Parameters	Values
0	μ_0	3
1	μ_1	4
2	σ_0	1
3	σ_1	3

$$\sum_{i=1}^{c-1} \eta_i(x) = 1$$

we can get that

$$\min \max \eta(X) = \frac{1}{c} \quad (56)$$

Combining Equation 55 and Equation 56 together, the maximum Bayes error is $1 - \frac{1}{c}$

Problem 2.7

Consider the following univariate Gaussian class-conditional densities:

$$p(x|Y=0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-3)^2}{2}\right) \quad (57)$$

$$p(x|Y=1) = \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{(x-4)^2}{18}\right) \quad (58)$$

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

The PDF of Guasssian distribution is¹

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

¹Gaussian PDF. WolframAlpha. URL: <https://mathworld.wolfram.com/GaussianFunction.html>

(a)

Draw the densities and determine the Bayes classifier graphically.

- The plot is displayed in Figure 1.
- The decision bounding was determined by the right intersection of both distributions. I applied Brent's method to find the intersection².
 - Intuitively, the intersection on the right has the minimum ϵ^0 to the right and ϵ^1 to the left.

```
1 class Gauss:
2     def __init__(self, scale, mean, var):
3         self.scale = scale
4         self.mean = mean
5         self.var = var
6     def pdf(self, x):
7         return 1/(self.scale*np.sqrt(2*np.pi))*np.exp(-1*(x-self.mean)**2/self.var)
8     def plot(self, ax, x_bound=[-5,13], nticks=200, **args):
9         xs = np.linspace(x_bound[0], x_bound[1], nticks)
10        ps = [self.pdf(x) for x in xs]
11        ax.plot(xs, ps, **args)
12
13
14 g0 = Gauss(1,3,2)
15 g1 = Gauss(3,4,18)
16
17 ## Boundaries
18 dec_x = [1.26, 4.49] # see problem 2.7 (b) for derivation
19
20 ## Plotting
21 fig, ax = plt.subplots()
22 g0.plot(ax, color="black",label="$p(x|Y=0) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-3)^2}{2})$")
23 g1.plot(ax, color="gray",linestyle="--",label="$p(x|Y=1)=\frac{1}{3\sqrt{2\pi}}\exp(-\frac{(x-4)^2}{18})$")
24 ax.axvline(x=dec_x[0], label="Bayes decision boundary")
25 ax.axvline(x=dec_x[1])
26 ax.set_xlabel("$x$")
27 ax.set_ylabel("$PDF$")
28 ax.annotate("$\psi^*(x)=1$", (7.5,0.2))
29 ax.annotate("$\psi^*(x)=1$", (-5,0.2))
30 ax.annotate("$\psi^*(x)=0$", (1.4,0.2))
```

²`scipy.optimize.brentq`. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.brentq.html>

```
31 ax.legend(loc="upper right");
```

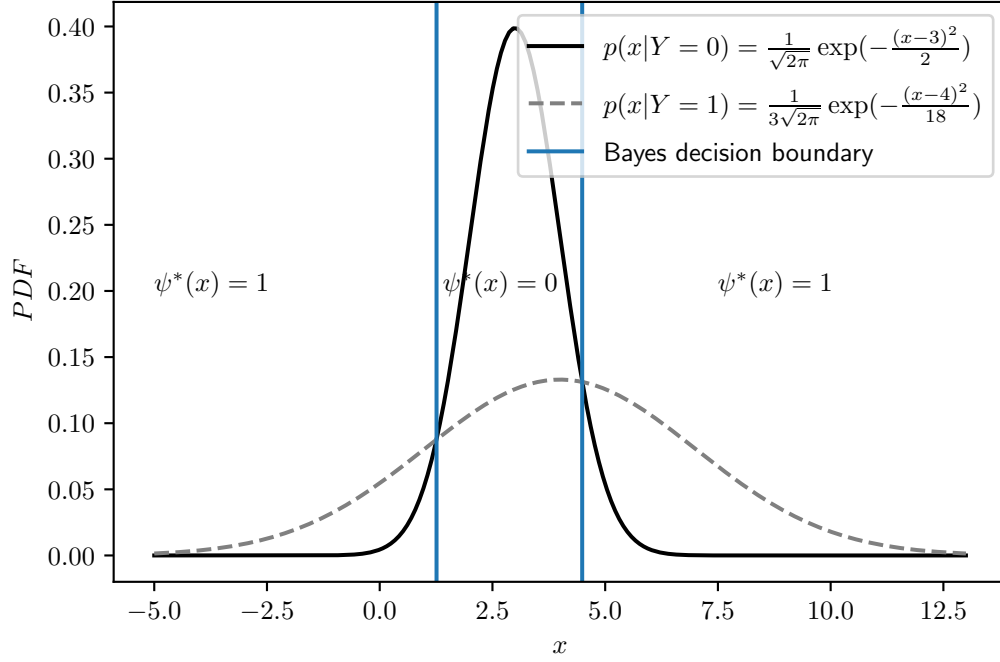


Figure 1: Univariate gaussian densities.

(b)

Determine the Bayes classifier.

According to Braga-Neto (2020, 22), the Bayes classifier can be defined by

$$\psi^*(x) = \begin{cases} 1, & D^*(x) > k^* \\ 0, & \text{otherwise} \end{cases} \quad (59)$$

where $D^*(x) = \ln \frac{p(x|Y=1)}{p(x|Y=0)}$, $k^* = \ln \frac{P(Y=0)}{P(Y=1)}$. Now, take Equation 57 and Equation 58 into the formula.

$$k^* = \ln \frac{1}{1} = 0$$

$$D^*(x) = \ln \frac{p(x|Y=1)}{p(x|Y=0)} \quad (60)$$

$$= \ln \frac{\frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-3)^2}{2})}{\frac{1}{3\sqrt{2\pi}} \exp(-\frac{(x-4)^2}{18})} \quad (61)$$

$$= \ln \left[3 \cdot \exp(-\frac{(x-3)^2}{2} + \frac{(x-4)^2}{18}) \right] \quad (62)$$

$$= \ln 3 - \frac{(x-3)^2}{2} + \frac{(x-4)^2}{18} \quad (63)$$

$$= \frac{-4}{9}x^2 + \frac{23}{9}x + (\ln(3) - \frac{65}{18}) \quad (64)$$

Thus, the Bayes classifier for distinguishing Equation 57 and Equation 58 is

$$\psi^*(x) = \begin{cases} 1, & \frac{-4}{9}x^2 + \frac{23}{9}x + (\ln(3) - \frac{65}{18}) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (65)$$

with the boundaries

$$x = \begin{cases} \frac{1}{8}(23 - 3\sqrt{1 + 16\ln(3)}) & \approx 1.26 \\ \frac{1}{8}(23 + 3\sqrt{1 + 16\ln(3)}) & \approx 4.49 \end{cases} \quad (66)$$

Noted that there are two boundaries for $D^*(x) = 0$ because $D^*(x)$ is a second order equation of x .

$$\psi^*(x) = \begin{cases} 1, & \left[x - (\frac{1}{8}(23 - 3\sqrt{1 + 16\ln(3)})) \right] \left[x - (\frac{1}{8}(23 + 3\sqrt{1 + 16\ln(3)})) \right] > 0 \\ 0, & \text{otherwise} \end{cases} \quad (67)$$

$$= \begin{cases} 1, & x < (\frac{1}{8}(23 - 3\sqrt{1 + 16\ln(3)})) \vee x > (\frac{1}{8}(23 + 3\sqrt{1 + 16\ln(3)})) \\ 0, & \text{otherwise} \end{cases} \quad (68)$$

$$\approx \begin{cases} 1, & x < 1.26 \vee x > 4.49 \\ 0, & \text{otherwise} \end{cases} \quad (69)$$

(c)

Determine the specificity and sensitivity of the Bayes classifier.

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

Let left and right boundaries be $\frac{1}{8}(23 - 3\sqrt{1 + 16\ln(3)}) = b_1$ and $\frac{1}{8}(23 + 3\sqrt{1 + 16\ln(3)}) = b_2$,

Table 2: The definition of sensitivity and specificity from Braga-Neto (2020, 18)

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

The standard normal CDF is³. Use the definition of $\epsilon^0[\psi]$ and $\epsilon^1[\psi]$ in Braga-Neto (2020, 18)

$$F(x) = p(X < x) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right) \right]$$

$$\epsilon^0[\psi^*(x)] = P(\psi(X) = 1 | Y = 0) \tag{70}$$

$$= \int_{\{x | \psi(x)=1\}} p(x | Y = 0) dx \tag{71}$$

$$= \int_{-\infty}^{b_1} p(x | Y = 0) dx + \int_{b_2}^{\infty} p(x | Y = 0) dx \tag{72}$$

$$= F_{X_0}(b_1) + 1 - F_{X_0}(b_2) \tag{73}$$

$$= \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{b_1 - \mu_0}{\sigma_0\sqrt{2}}\right) \right] + 1 - \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{b_2 - \mu_0}{\sigma_0\sqrt{2}}\right) \right] \tag{74}$$

$$= 1 + \frac{1}{2} \left(\operatorname{erf}\left(\frac{b_1 - \mu_0}{\sigma_0\sqrt{2}}\right) - \operatorname{erf}\left(\frac{b_2 - \mu_0}{\sigma_0\sqrt{2}}\right) \right) \tag{75}$$

³Normal distribution. Wiki URL: https://en.wikipedia.org/wiki/Normal_distribution

Table 3: Exact values of type 0 and type 1 error rates.

	Error statistics	Value
0	epsilon_0	0.108752
1	epsilon_1	0.384628
2	Sensitivity	0.615372
3	Specificity	0.891248
4	Overall Bayes error	0.246690

$$\epsilon^1[\psi^*(x)] = P(\psi(X) = 0|Y = 1) \quad (76)$$

$$= \int_{\{x|\psi(x)=0\}} p(x|Y=1)dx \quad (77)$$

$$= \int_{b_1}^{b_2} p(x|Y=1)dx \quad (78)$$

$$= F_{X_1}(b_2) - F_{X_1}(b_1) \quad (79)$$

$$= \frac{1}{2} \left[\operatorname{erf}\left(\frac{b_2 - \mu_1}{\sigma_1 \sqrt{2}}\right) - \operatorname{erf}\left(\frac{b_1 - \mu_1}{\sigma_1 \sqrt{2}}\right) \right] \quad (80)$$

$$(81)$$

The exact values of error estimation are shown in Table 3 (epsilon_0 is ϵ_0 , and epsilon_1 ϵ_1). Specificity and sensitivity are calculated with their definitions shown in Table 2,

```

1 # Calculation of error statistics
2 def nerf(b, mu, std):
3     return erf( (b - mu) / (std*np.sqrt(2)))
4
5 b1 = (1/8)*(23-3*np.sqrt(1+16*np.log(3)))
6 b2 = (1/8)*(23+3*np.sqrt(1+16*np.log(3)))
7 e0 = 1 + 0.5*(nerf(b1,mu0, std0) - nerf(b2,mu0, std0))
8 e1 = 0.5*(nerf(b2,mu1, std1) - nerf(b1, mu1, std1))
9
10 sensi = 1 - e1
11 spec = 1 - e0
12 bayesError = 0.5*(e0+e1)

```


Table 4: Exact values of Bayes error.

Error statistics	Value
0 Overall Bayes error	0.24669

(d)

Determine the overall Bayes error.

Use the derivation in Braga-Neto (2020, 18),

$$\epsilon[\psi^*(X)] = P(\psi(X) \neq Y) \quad (82)$$

$$= P(\psi(X) = 1, Y = 0) + P(\psi(X) = 0, Y = 1) \quad (83)$$

$$= P(Y = 0)P(\psi(X) = 1|Y = 0) + P(Y = 1)P(\psi(X) = 0|Y = 1) \quad (84)$$

$$= P(Y = 0)\epsilon^0 + P(Y = 1)\epsilon^1 \quad (85)$$

$$= \frac{1}{2}(\epsilon^0 + \epsilon^1) \quad (86)$$

The exact Bayes error is displayed in Table 4.

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.

Since $\Sigma_0 \neq \Sigma_1$ happens in the following subproblems, these are *heteroskedastic cases*. As mentioned in Braga-Neto (2020, sec. 2.5.2). The Bayes classifier is

$$\psi_Q^*(x) = \begin{cases} 1, & x^T A x + b^T x + c > 0 \\ 0, & \text{otherwise} \end{cases} \quad (87)$$

where

$$A = \frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1}) \quad (88)$$

$$b = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0 \quad (89)$$

$$c = \frac{1}{2}(\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1) + \frac{1}{2} \ln \frac{\det(\Sigma_0)}{\det(\Sigma_1)} + \ln \frac{P(Y=1)}{P(Y=0)} \quad (90)$$

Let x be the vector of sample values,

$$x = \begin{bmatrix} x_1 \\ x_0 \end{bmatrix} \quad (91)$$

The following is the Python implementation with NumPy^{4 5}.

```

1  def compose(U,D):
2      return U.transpose()@inv(D)@U
3
4  def mA(sigma0, sigma1):
5      sigma0_inv = inv(sigma0)
6      sigma1_inv = inv(sigma1)
7      return 0.5*(sigma0_inv - sigma1_inv)
8
9  def mb(sigma0, sigma1, mu0, mu1):
10     sigma0_inv = inv(sigma0)
11     sigma1_inv = inv(sigma1)
12     return sigma1_inv@mu1 - sigma0_inv@mu0
13
14  def mc(sigma0, sigma1, mu0, mu1, Py):
15     return 0.5*(compose(mu0, sigma0) - compose(mu1, sigma1)) +\
16         0.5*np.log(det(sigma0)/det(sigma1)) +\
17         np.log(Py/(1-Py))
18
19  def BayesBound(x, sigma0, sigma1, mu0, mu1, Py):
20     xax = x.transpose() @ mA(sigma0, sigma1)@x
21     bx = mb(sigma0, sigma1, mu0, mu1).transpose() @ x
22     c = mc(sigma0, sigma1, mu0, mu1, Py)
23

```

⁴There is a well-written documentation for matrix operation: https://numpy.org/doc/stable/user/absolute_beginners.html#creating-matrices

⁵the contour plot for Gaussian process is referred to <https://gist.github.com/gwgundersen/90dfa64ca29aa8c3833dbc6b03de44be>.

```

24     return float(xax + bx + c)
25
26 class GaussianBayesClassifier:
27     def __init__(self, sigma0, sigma1, mu0, mu1, Py):
28         self.sigma0 = sigma0
29         self.sigma1 = sigma1
30         self.mu0 = mu0
31         self.mu1 = mu1
32         self.Py = Py
33
34         # Inferred Matrix
35         self.mA = mA(sigma0, sigma1)
36         self.mb = mb(sigma0, sigma1, mu0, mu1)
37         self.mc = mc(sigma0, sigma1, mu0, mu1, Py)
38
39     def BayesBound(self, x):
40         return BayesBound(x, self.sigma0, self.sigma1, self.mu0, self.mu1, self.Py)
41
42     def psi(self, x):
43         """
44         Bayes classification
45         """
46         pred = 0
47         if self.BayesBound(x) > 0:
48             pred = 1
49         return pred
50
51     def plot2D(self, psi_annotates=[[0.3,0.3], [0.7,0.7]]):
52         fig, ax = plt.subplots()
53
54         # Create grids
55         xlist = np.linspace(-3.5,3.5,RES_GRID)
56         ylist = np.linspace(-3.5,3.5,RES_GRID)
57         X, Y = np.meshgrid(xlist, ylist)
58         pos = np.dstack((X,Y))
59
60         # Compute Bayes classification
61         Z = np.zeros(X.shape)
62         for i in range(0, Z.shape[0]):
63             for j in range(0, Z.shape[1]):
64                 x = np.matrix([X[i,j], Y[i,j]]).T

```

```

65         Z[i,j] = self.psi(x)
66
67     # Compute Gaussia pdf
68     rv0 = multivariate_normal(np.array(self.mu0.T)[0], self.sigma0)
69     rv1 = multivariate_normal(np.array(self.mu1.T)[0], self.sigma1)
70     Z0 = rv0.pdf(pos)
71     Z1 = rv1.pdf(pos)
72
73     # Plot contours
74     cmap = plt.get_cmap('Set1', 2)
75     ax.contour(X,Y,Z, cmap=cmap, levels=[0.9])
76     ax.contour(X,Y,Z0, alpha=0.4)
77     ax.contour(X,Y,Z1, alpha=0.4)
78     ax.plot(self.mu0[0,0], self.mu0[1,0], marker="o", color="k",label="$\mu_0$")
79     ax.plot(self.mu1[0,0], self.mu1[1,0], marker="o", color="gray",label="$\mu_1$")
80     if psi_annotates==None:
81         psi_annotates = [[self.mu0[0,0], self.mu0[1,0]], [self.mu1[0,0], self.mu1[1,0]]]
82
83     # Annotate decisions
84     for ann in psi_annotates:
85         i = X[int(X.shape[0]*ann[0]), int(X.shape[1]*ann[1])]
86         j = Y[int(Y.shape[0]*ann[0]), int(Y.shape[1]*ann[1])]
87         x = np.matrix([i, j]).T
88         if self.psi(x) > 0:
89             ax.annotate("$\psi^{*}(x) = 1$", (i, j))
90         else:
91             ax.annotate("$\psi^{*}(x) = 0$", (i, j))
92
93     # legend and label settings
94     ax.set_xlabel("$x_1$")
95     ax.set_ylabel("$x_2$")
96     ax.legend()

```

(a)

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

```

1 bc = GaussianBayesClassifier(np.matrix([[2,0],[0,1]]),\
2                               np.matrix([[2,0],[0,4]]),\
3                               np.matrix([0,0]).T,\
4                               np.matrix([2,0]).T, 0.5)
5 bc.plot2D()

```

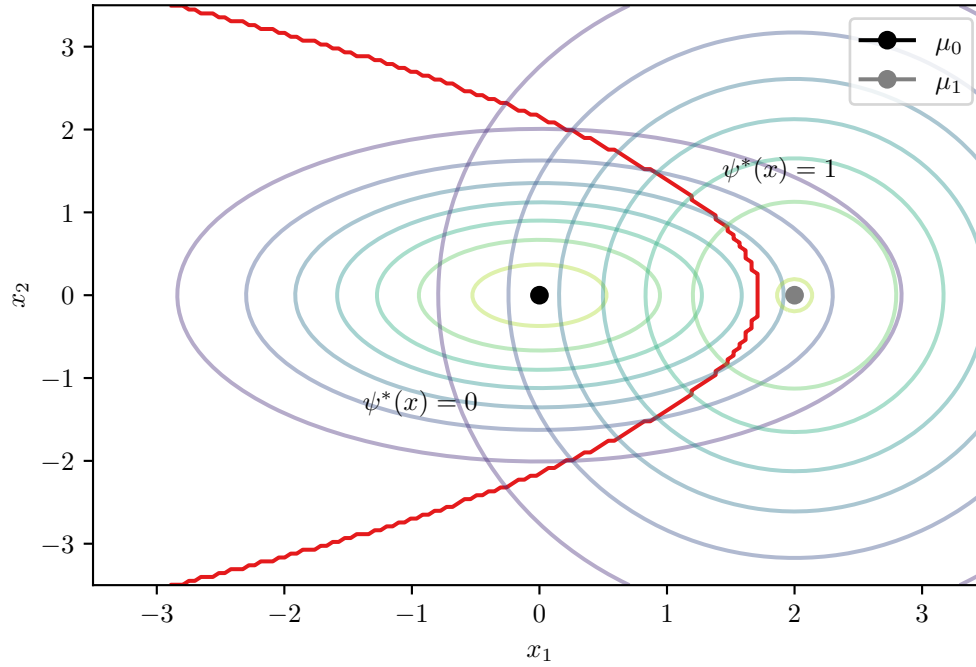


Figure 2: Bayes Classifier for 2D Gaussian (a)

(b)

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

```

1 bc = GaussianBayesClassifier(np.matrix([[2,0],[0,1]]),\
2                               np.matrix([[4,0],[0,1]]),\
3                               np.matrix([0,0]).T,np.matrix([2,0]).T,\
4                               0.5)
5 bc.plot2D()

```

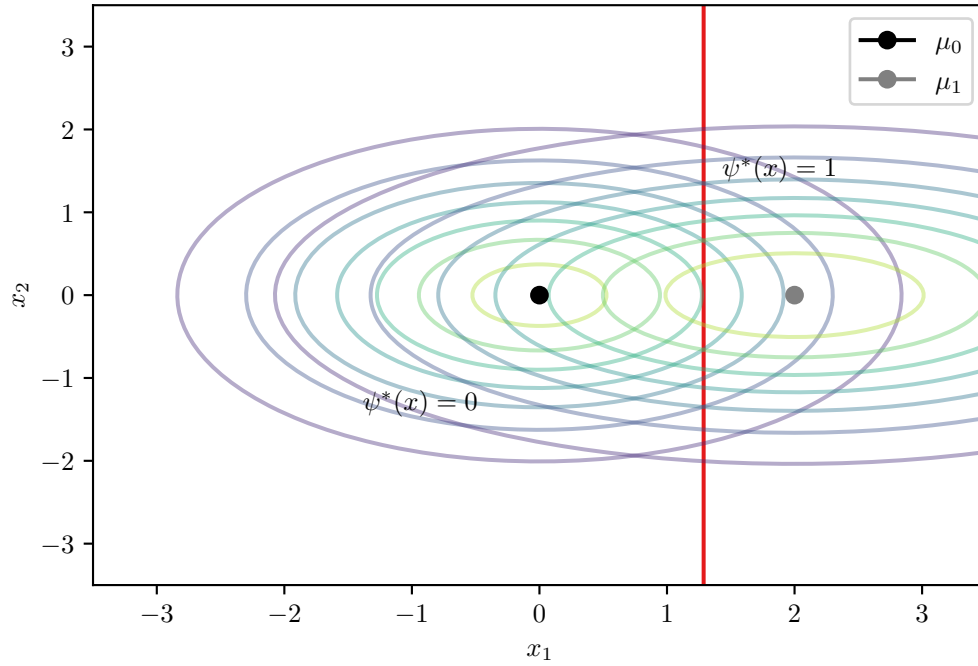


Figure 3: Bayes Classifier for 2D Gaussian (b)

(c)

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

```

1 bc = GaussianBayesClassifier(np.matrix([[1,0],[0,1]]),\
2                               np.matrix([[2,0],[0,2]]),\
3                               np.matrix([0,0]).T,\
4                               np.matrix([0,0]).T,\
5                               0.5)
6 bc.plot2D(psi_annotates= [[0.4,0.3], [0.7,0.7]])

```

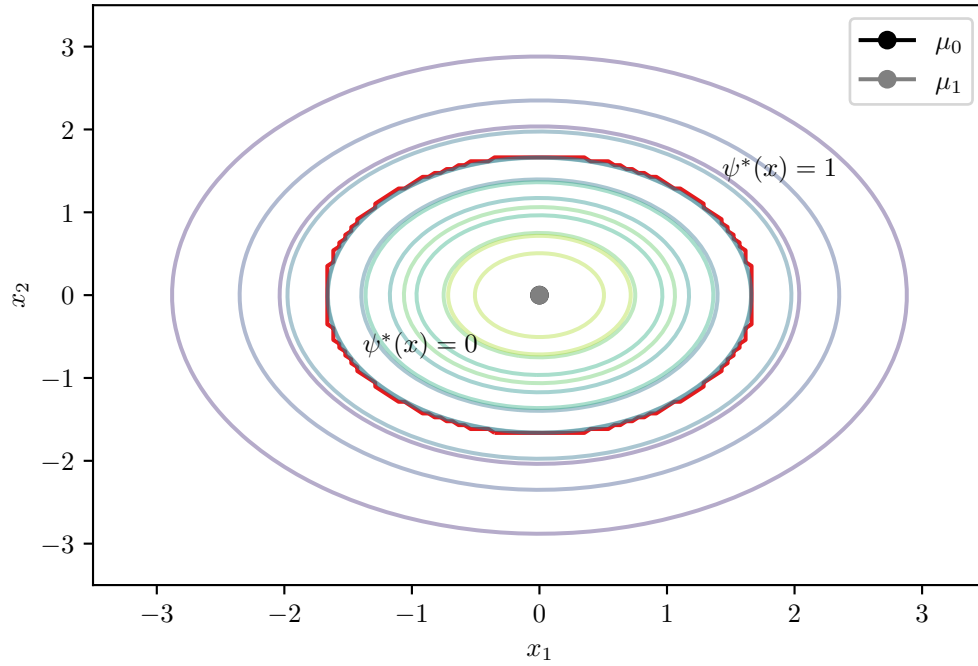


Figure 4: Bayes Classifier for 2D Gaussian (c)

(d)

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

```

1 bc = GaussianBayesClassifier(np.matrix([[2,0],[0,1]]),\
2                               np.matrix([[1,0],[0,2]]),\
3                               np.matrix([0,0]).T,\
4                               np.matrix([0,0]).T,\
5                               0.5)
6 bc.plot2D(psi_annotates= [[0.3,0.4], [0.6,0.7]])

```

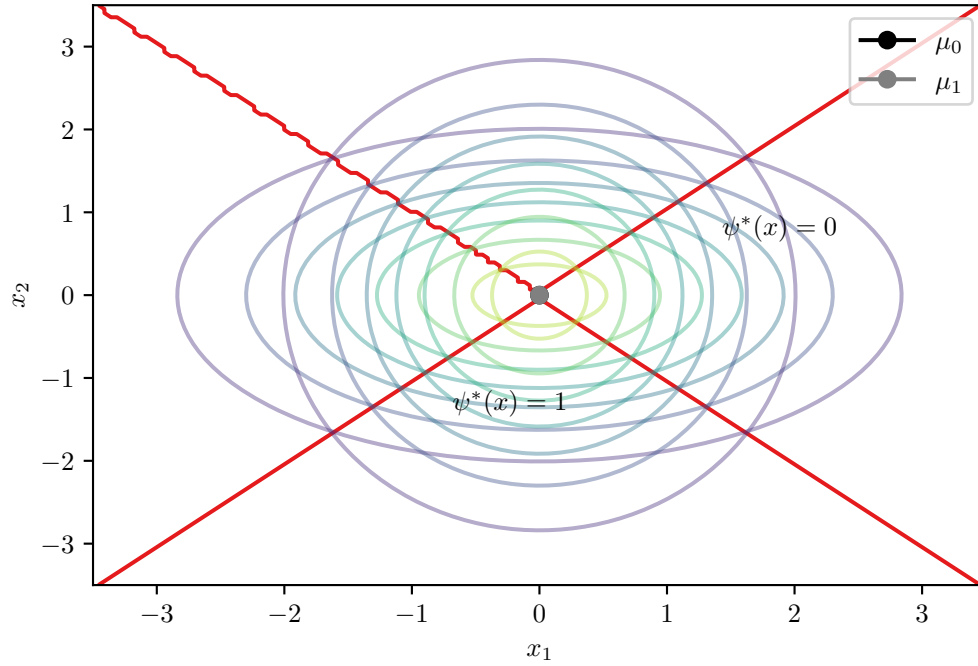


Figure 5: Bayes Classifier for 2D Gaussian (d)

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 ρ_1, \dots, ρ_k .

Hint: Use the fact that

$$\begin{bmatrix} 1 & \sigma & \cdots & \sigma \\ \sigma & 1 & \cdots & \sigma \\ \vdots & \vdots & \ddots & \vdots \\ \sigma & \sigma & \cdots & 1 \end{bmatrix}_{l \times l}^{-1} = \frac{1}{(1-\sigma)(1+(l-1)\sigma)} \begin{bmatrix} 1+(l-2)\sigma & -\sigma \cdots & -\sigma & \\ -\sigma & 1+(l-2)\sigma & \cdots & -\sigma \\ \vdots & \vdots & \ddots & \vdots \\ -\sigma & -\sigma & \cdots & 1+(l-2)\sigma \end{bmatrix} \sigma \quad (92)$$

Use Equation 2.53 in Braga-Neto (2020, 31). Given $P(Y = 0) = P(Y = 1) = 0.5$, then

$$\epsilon_L^* = \Phi\left(-\frac{\delta}{2}\right) \quad (93)$$

where $\delta = \sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$.

$$\Sigma_{d \times d} = \begin{bmatrix} \Sigma_{l_1 \times l_1}(\sigma_1^2, \rho_1) & 0 & \cdots & 0 \\ 0 & \Sigma_{l_2 \times l_2}(\sigma_2^2, \rho_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{l_k \times l_k}(\sigma_k^2, \rho_k) \end{bmatrix} \quad (94)$$

where $l_1 + \cdots + l_k = d$.

$$\Sigma_{l_k \times l_k}(\sigma_i^2, \rho_i) = \sigma_i^2 \begin{bmatrix} 1 & \rho_i & \cdots & \rho_i \\ \rho_i & 1 & \cdots & \rho_i \\ \vdots & \vdots & \ddots & \vdots \\ \rho_i & \rho_i & \cdots & 1 \end{bmatrix} \quad (95)$$

$$\Sigma_{l_k \times l_k}^{-1}(\sigma_i^2, \rho_i) = \frac{1}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} \begin{bmatrix} 1+(l_k-2)\rho_i & -\rho_i & \cdots & -\rho_i \\ -\rho_i & 1+(l_k-2)\rho_i & \cdots & -\rho_i \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_i & -\rho_i & \cdots & 1+(l_k-2)\rho_i \end{bmatrix} \quad (96)$$

$$= \begin{bmatrix} \frac{1+(l_k-2)\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} & \frac{-\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} & \cdots & \frac{-\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} \\ \frac{-\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} & \frac{1+(l_k-2)\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} & \cdots & \frac{-\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} & \cdots & \frac{-\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} & \frac{1+(l_k-2)\rho_i}{\sigma_i^2(1-\rho_i)(1+(l_k-1)\rho_i)} \end{bmatrix} \quad (97)$$

The element-wised summation of all terms in $\Sigma_{l_k \times l_k}(\sigma_i^2, \rho_i)$ is

$$\text{sum}(\Sigma_{l_k \times l_k}(\sigma_k^2, \rho_k)) = \frac{1}{\sigma_k^2(1 - \rho_k)(1 + (l_k - 1)\rho_k)}(-l_k^2\rho_k + l_k(1 + (l_k - 1)\rho_k)) \quad (98)$$

$$= \frac{l_k(1 - \rho_k)}{\sigma_k^2(1 - \rho_k)(1 + (l_k - 1)\rho_k)} \quad (99)$$

Thus, the inverse of covariance matrix is

$$\Sigma_{d \times d}^{-1} = \begin{bmatrix} \Sigma_{l_1 \times l_1}^{-1}(\sigma_1^2, \rho_1) & 0 & \cdots & 0 \\ 0 & \Sigma_{l_2 \times l_2}^{-1}(\sigma_2^2, \rho_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{l_k \times l_k}^{-1}(\sigma_k^2, \rho_k) \end{bmatrix} \quad (100)$$

Combining together,

$$\epsilon_L^* = \Phi\left(\frac{-1}{2}\delta\right) = \Phi\left(\frac{-1}{2}\sqrt{(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0)}\right) \quad (101)$$

$$= \Phi\left(\frac{-1}{2}\sqrt{\underbrace{\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}_{1 \times d} \begin{bmatrix} \Sigma_{l_1 \times l_1}^{-1}(\sigma_1^2, \rho_1) & 0 & \cdots & 0 \\ 0 & \Sigma_{l_2 \times l_2}^{-1}(\sigma_2^2, \rho_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{l_k \times l_k}^{-1}(\sigma_k^2, \rho_k) \end{bmatrix}_{d \times d} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{d \times 1}}}_{d \times d}}\right) \quad (102)$$

$$= \Phi\left(\sum_{i=1}^k \frac{l_i(1 - \rho_i)}{\sigma_i^2(1 - \rho_i)(1 + (l_i - 1)\rho_i)}\right) \quad (103)$$

(b)

Specialize the previous formula for equal-sized blocks $l_1 = \cdots = l_k = l$ with equal correlations $\rho_1 = \cdots = \rho_k = \rho$, and constant variance $\sigma_1^2 = \cdots, \sigma_k^2 = \sigma^2$. Write the resulting formula in terms of d, l, σ and ρ .

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 $\rho = 0, 0.25, 0.5, 0.75$ (plot one curve for each value). Confirm that the Bayes error increases monotonically with σ from 0 to 0.5 for each value of ρ , and that

Bayes error for larger ρ is uniformly larger than that for smaller ρ . The latter fact shows that correlation between the features is detrimental to classification.

As shown in Figure 6, the correlation does have the detrimental effect on classification, and monotoniously increases Bayes error.

```

1  # Implementation
2  def get_VecM(length, val):
3      return np.ones((1, length)).T * val
4
5  class GeneralMultiGaussian:
6      def __init__(self, d, l, sig, rho, Imu0=0, Imu1=1):
7          self.d = d
8          self.l = l
9          self.sig = sig
10         self.rho = rho
11         self.Imu0 = Imu0
12         self.Imu1 = Imu1
13         self.mu0 = get_VecM(d, Imu0)
14         self.mu1 = get_VecM(d, Imu1)
15         # cluster of covariance matrix
16         self.subCovInv = self.cluster_cov_inv()
17         self.CovInv = self.cov_inv()
18
19     def cluster_cov_inv(self):
20         sig = self.sig
21         l = self.l
22         rho = self.rho
23         scale = 1/((sig**2)*(1-rho)*(1+(l-1)*rho))
24         diag_val = 1 + (l-2)*rho
25         covInv = np.ones((l,l)) * (-1*rho)
26         covInv = np.matrix(covInv)
27         np.fill_diagonal(covInv, diag_val)
28         return scale*covInv
29
30     def cov_inv(self):
31         d = self.d; l = self.l
32         subMs = [self.subCovInv for i in range(0,int(d/l)+1)]
33         return block_diag(*subMs)[0:d, 0:d]
34
35     def bayesError(self):
36         mu0 = self.mu0

```

```

37         mu1 = self.mu1
38         mud = mu1 - mu0
39         CovInv = self.CovInv
40         delta = -0.5 * np.sqrt(mud.T @ CovInv @ mud)
41         return norm.cdf(delta)
42
43
44 # Parameter Setting
45 sigs = np.linspace(0.01 ,3 , 50)
46 d = 20
47 l = 4
48 rhos = [0, 0.25, 0.5, 0.75]
49
50 # Measurement of Bayes errors
51 errorD = dict()
52 for rho in rhos:
53     errorD[rho] = np.zeros(len(sigs))
54
55 for rho in errorD.keys():
56     for (i, sig) in enumerate(sigs):
57         gm = GeneralMultiGaussian(d, l, sig, rho)
58         err = gm.bayesError()
59         errorD[rho][i] = err
60
61 # Plotting
62 fig, ax = plt.subplots()
63 for rho in rhos:
64     ax.plot(sigs, errorD[rho], label="$\\rho={}$".format(rho))
65 ax.set_xlabel("$\\sigma$")
66 ax.set_ylabel("Bayes Error")
67 ax.legend();

```

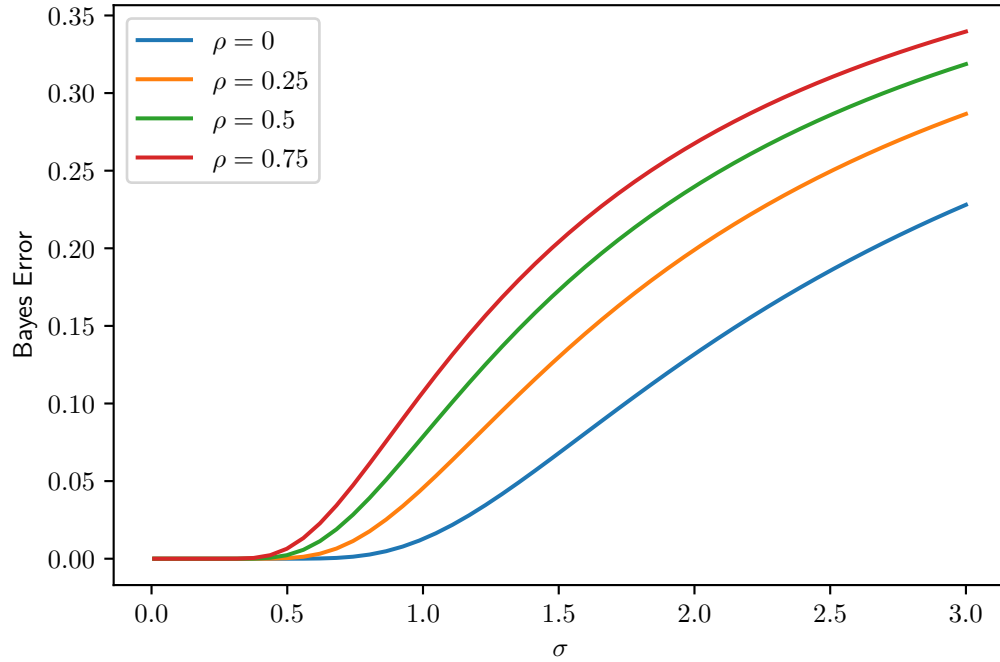


Figure 6: The relation of Bayes error with standard deviation and covariance.

ii.

Plot the Bayes error as a function of $d = 2, 4, 6, 8, \dots, 40$, with fixed block size $l = 4$ and variance $\sigma^2 = 1$ and $\rho = 0, 0.25, 0.5, 0.75$ (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.

The plot is shown in Figure 7.

```

1 # Parameter setting
2 ds = np.arange(4, 40, 4, dtype=int)
3 l = 4
4 sig = 1
5
6 # Measure errors
7 errorD2 = dict()
8 for rho in rhos:
9     errorD2[rho] = np.zeros(len(ds))
10
11 for rho in errorD2.keys():

```

```

12     for (i, d) in enumerate(ds):
13         gm = GeneralMultiGaussian(d, l, sig, rho)
14         err = gm.bayesError()
15         errorD2[rho][i] = err
16
17 # Plotting
18 fig, ax = plt.subplots()
19 for rho in rhos:
20     ax.plot(ds, errorD2[rho], label="$\\rho={}$".format(rho))
21 ax.set_xlabel("$d$")
22 ax.set_ylabel("Bayes Error")
23 ax.legend();

```

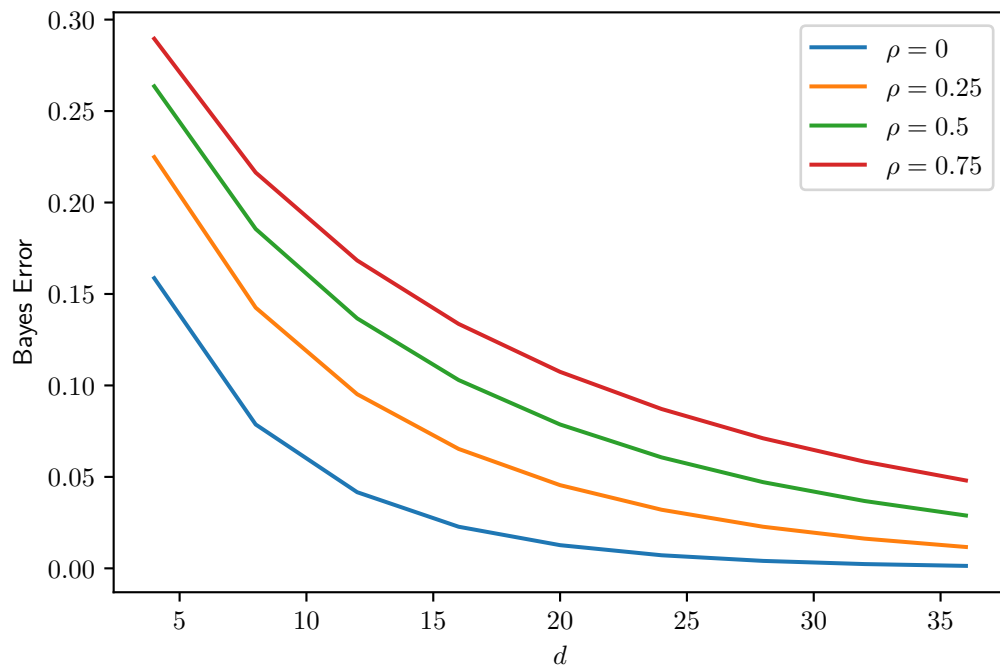


Figure 7: Bayes error as a function of dimension.

iii.

Plot the Bayes error as a function of the correlation $\rho \in [0, 1]$ 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 $\rho = 0$, which

shows that the Bayes error is very sensitive to correlation in the near-independent region.

The plot is show in Figure 8.

```
1 # Parameter setting
2 d = 20
3 ls = [1,2,4,10]
4 sig = 2
5 rhos = np.linspace(0.,0.99,40)
6
7 # Measure errors
8 errorD3 = dict()
9 for l in ls:
10     errorD3[l] = np.zeros(len(rhos))
11
12 for l in ls:
13     for (i, rho) in enumerate(rhos):
14         gm = GeneralMultiGaussian(d, l, sig, rho)
15         err = gm.bayesError()
16         errorD3[l][i] = err
17
18 # Plotting
19 fig, ax = plt.subplots()
20 for l in ls:
21     ax.plot(rhos, errorD3[l], label="$\\l={}$".format(l))
22 ax.set_xlabel("$\\rho$")
23 ax.set_ylabel("Bayes Error")
24 ax.legend();
```

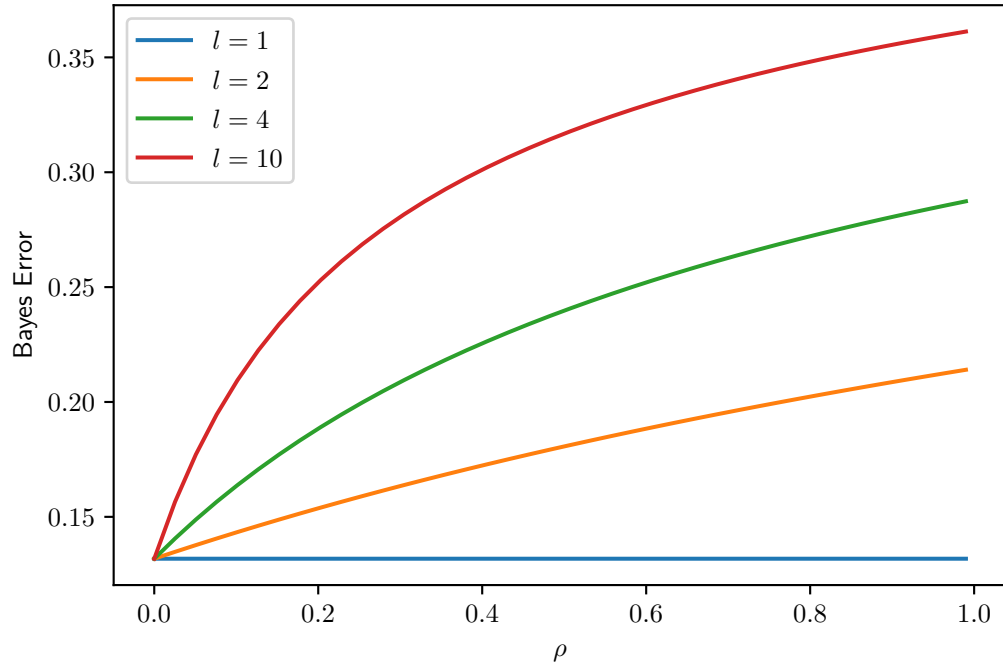


Figure 8: Bayes error as a function of correlation.

References

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