

**Office of Sustainability**  
Amirkabir University of Technology  
(Tehran Polytechnic)

COMPUTER ENGINEERING && IT DEPARTMENT  
AMIRKABIR UNIVERSITY OF TECHNOLOGY

---

# Statistical Pattern Recognition

---

*Submitted To:*  
Mohammad Rahmati  
Asst. Professor  
Computer Engineering  
Department

*Submitted By :*  
Ahmad Asadi  
94131091  
Group-G1  
Fall-95

## Contents

1	Problem 1	2
2	Problem 2	5
3	Problem 3	7
4	Problem 4	9
5	Problem 5	12
6	Problem 6	14
7	Problem 7	16
8	Problem 8	17
9	Problem 9	20
10	Problem 10	21
11	Problem 11	22

# 1 Problem 1

Two normal distribution are characterized by:

$$P_1 = P_2 = 0.5$$

$$M_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

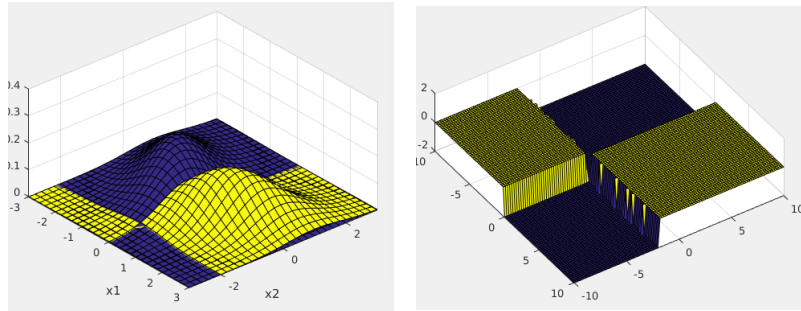
$$M_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

1. Draw the Bayes decision boundary to minimize the probability of error.

The bayesian decision function is:

$$\frac{P(X|w_1) \cdot P(w_1)}{P(X|w_2) \cdot P(w_2)} \underset{<_{w_2}}{\overset{>_{w_1}}{1}} \rightarrow \frac{P(X|w_1)}{P(X|w_2)} \underset{<_{w_2}}{\overset{>_{w_1}}{1}}$$

The figure 1 displays the data distributions and decision boundaries of bayesian decision function without considering costs.



(a) Distribution of data. The distribution of class 1 is displayed in blue color and the distribution of class 2 is displayed in yellow. (b) Decision Boundary. The decision boundary of class 1 is displayed in blue color and the decision boundary of class 2 is displayed in yellow.

Figure 1: Bayesian Decision Boundary without considering decision costs

Also, it is possible to transform the data to a one dimensional space first and make decision in the new space using a single threshold as bellow:

$$\begin{aligned}
h(X) &= \frac{1}{2}(X - M_1)^T \Sigma_1^{-1}(X - M_1) - \frac{1}{2}(X - M_2)^T \Sigma_2^{-1}(X - M_2) \\
&\quad + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} \ln \frac{P(W_1)}{P(W_2)} \\
&\rightarrow h(X) = \frac{1}{2}(X - \begin{bmatrix} 1 \\ 0 \end{bmatrix})^T \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}^{-1} (X - \begin{bmatrix} 1 \\ 0 \end{bmatrix}) \\
&\quad - \frac{1}{2}(X - \begin{bmatrix} -1 \\ 0 \end{bmatrix})^T \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}^{-1} (X - \begin{bmatrix} -1 \\ 0 \end{bmatrix}) \\
&\quad + \frac{1}{2} \ln \frac{\left| \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right|}{\left| \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \right|} \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} 0 \\
&\rightarrow h(X) = \frac{1}{2}(X - \begin{bmatrix} 1 \\ 0 \end{bmatrix})^T \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}^{-1} (X - \begin{bmatrix} 1 \\ 0 \end{bmatrix}) \\
&\quad - \frac{1}{2}(X - \begin{bmatrix} -1 \\ 0 \end{bmatrix})^T \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}^{-1} (X - \begin{bmatrix} -1 \\ 0 \end{bmatrix}) \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} 0
\end{aligned}$$


---

2. Draw the Bayes decision boundary to minimize the cost with  $c_{11} = c_{22} = 0$  and  $c_{12} = 2c_{21}$

The Bayesian decision boundary when considering costs of decisions turns to the form bellow:

$$\frac{p_1(X)}{p_2(X)} \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} \frac{(c_{12} - c_{22})P_2}{(c_{21} - c_{11})P_1} \quad (1)$$

Figure 2 displays the decision boundary in this case.

---

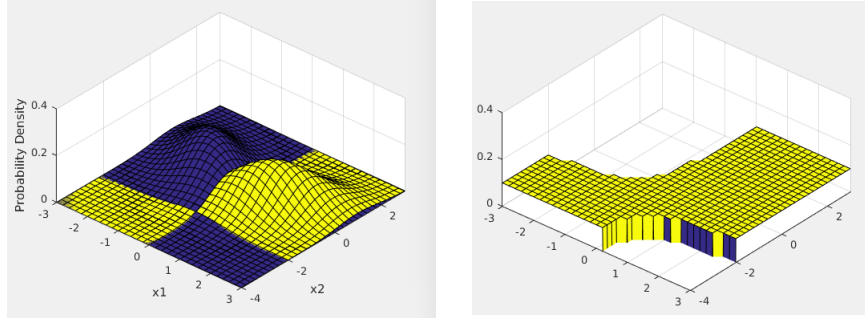
3. Assume that  $c_{11} = c_{22} = 0$  and  $c_{12} = c_{21}$  :

- (a) Plot the operating characteristics.

As equation (1) yields, with  $c_{11} = c_{22} = 0$  and  $c_{12} = c_{21}$  the decision boundary is not different from that of section 1. So the figure 1 is displaying the required decision boundary

---

- (b) Find the total error when Neyman-Pearson test is performed with  $\epsilon_1 = 0.05$



(a) Distribution of data. The distribution of class 1 is displayed in yellow color and the distribution of class 2 is displayed in blue.

(b) Decision Boundary. The decision boundary of class 1 is displayed in yellow color and the that of class 2 is displayed in blue.

Figure 2: Bayesian Decision Boundary considering decision costs  $c_{11} = c_{22} = 0$  and  $c_{12} = 2c_{21}$

We should estimate  $\mu$  such that the following function  $r$  be minimized.

$$r = \mu(\epsilon_1 - 0.05) + \epsilon_2 \quad (2)$$

After finding appropriate  $\mu$ , the decision boundary is derivable from:

$$\frac{p_1(X)}{p_2(X)} \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} \mu \quad (3)$$

To find such  $\mu$  an script will iterate on following equation to reach the best fitting value.

$$\epsilon_1 = \int_{\mu}^{\infty} p_h(h|W-1)dh = 0.05 \quad (4)$$

The  $\mu = 1.0795e - 78$  is the best value estimated with  $\epsilon_1 = 0.051$  error rate. Figure 3 illustrates the resulted decision boundary.

---

(c) Find the threshold value and total error for the minimax test.

---

(d) Plot the error-reject curve.

---

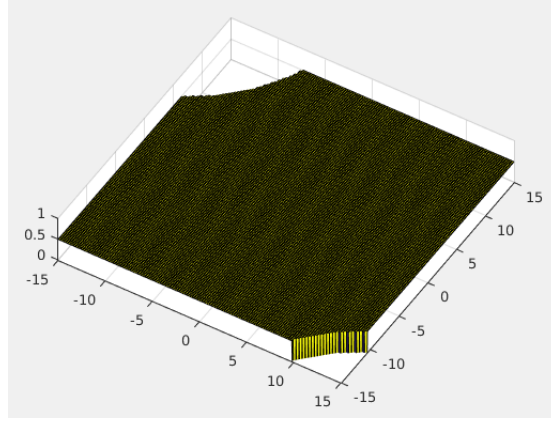


Figure 3: Decision boundary resulted from Neyman-Pearson test with  $\epsilon_1 = 0.05$

## 2 Problem 2

Two normal distribution are characterized by:

$$\begin{aligned}
 P_1 &= P_2 = 0.5 \\
 M_1 &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 4 & -3 \\ -3 & 4 \end{bmatrix} \\
 M_2 &= \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}
 \end{aligned}$$

1. Find the linear discriminant function which maximize the Fisher criterion and minimize the error by adjusting the threshold.

$$\begin{aligned}
 f &= \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \rightarrow \frac{\partial f}{\partial \sigma_1^2} = \frac{\partial f}{\partial \sigma_2^2} = \frac{-(\mu_1 - \mu_2)^2}{(\sigma_1^2 + \sigma_2^2)^2} \\
 &\quad \frac{\partial f}{\partial \sigma_1^2} \\
 \rightarrow s &= \frac{\frac{\partial f}{\partial \sigma_1^2}}{\frac{\partial f}{\partial \sigma_1^2} + \frac{\partial f}{\partial \sigma_2^2}} = 0.5 \\
 \rightarrow V &= [0.5\Sigma_1 + 0.5\Sigma_2]^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}^{-1} \begin{bmatrix} -2 \\ 0 \end{bmatrix} = \frac{1}{16} \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} -2 \\ 0 \end{bmatrix} \\
 \rightarrow V &= \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \tag{5}
 \end{aligned}$$

Equation (5) illustrates the best mapping vector  $V$  to project distribution over, therefore the LDA will be in the form of  $V^t x \geq \alpha$ , in which  $\alpha$  should be find

from equation (6).

$$\frac{\partial f}{\partial \mu_1} + \frac{\partial f}{\partial \mu_2} = 0 \quad (6)$$

which results:

$$\frac{2\mu_1 - 2\mu_2}{\sigma_1^2 + \sigma_2^2} + \frac{-2\mu_1 + 2\mu_2}{\sigma_1^2 + \sigma_2^2} = 0 \rightarrow 0 = 0$$

Therefore, the best  $\alpha$  is not derivable analytical. In order to estimate the best  $\alpha$  a MATLAB script using bootstrapping has been written by me, which is appended to the homework submission zip file. The bootstrapping process, iterates 1000 number of times in each iteration generating 1000 multivariate random variables regarding given means and variances for each class. We attempt to estimate best  $\alpha$  in each iteration using grid search in the range of  $[-4, 4]$  with step length 0.01. To measure the accuracy of chosen  $\alpha$  we use the ratio  $\frac{tp+tn-fp-fn}{N}$  as the measurement function. After all iterations, the mean of all best estimated  $\alpha$ s is reported as the best  $\alpha$ . Experiments resulted the  $\alpha = 0.23$  as the best value separating two classes from each other. So the LDA function is being reformed as bellow:

$$\begin{bmatrix} -0.5 \\ 0 \end{bmatrix}^T X \geq^{w_1} 0.23 \quad (7)$$

- 
2. Find the optimum linear discriminant function which minimize the probability of error.

The iterative process has been coded in MATLAB and its code is attached to the submitted homework zip file.

The parameter  $s$  has been estimated using a grid search in range  $[0, 1]$  with step length 0.01. the error probability *epsilon* in each iteration with respect to selected  $s$  has been calculated. The figure 4 illustrates the behaviour of error probability w.r.t. selected  $s$ . The best value of  $s$  is in 0 or 1.

To through a light over reality, the projection of data in two cases of  $s = 0.5$  and  $s = 1$  has been displayed in the figure 5, respectively the worst- and the best-case. As its obviously clear, the separation of classes is more suitable in case  $s = 1$  rather than  $s = 0.5$ , hence the lower error probability will be gained when  $s = 1$ .

---

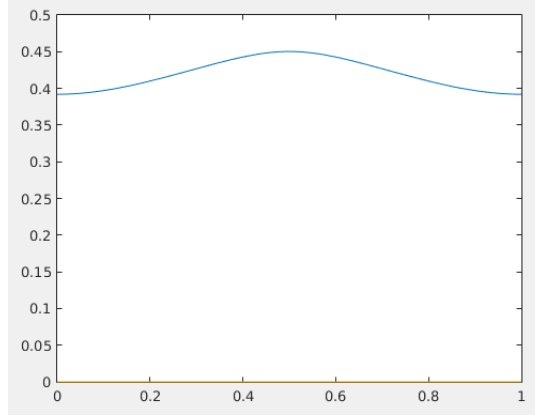


Figure 4: The behaviour of error probability  $\epsilon$  w.r.t. chosen  $s$

### 3 Problem 3

We consider a classification problem in dimension  $d = 2$ , with  $k = 3$  classes where:  $p(x|w_i) \propto N(\mu_i, \Sigma_i)$ ,  $i = 1, 2, 3$ , with:

$$\mu_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \mu_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_i = \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

1. Calculate the discriminant function  $g_i(x)$  for each class.

Generating a discriminant function between all pairs of classes can be a good solution. Assuming  $P(W_1) = P(W_2) = P(W_3)$ , the Bayes classifier is a good choice:

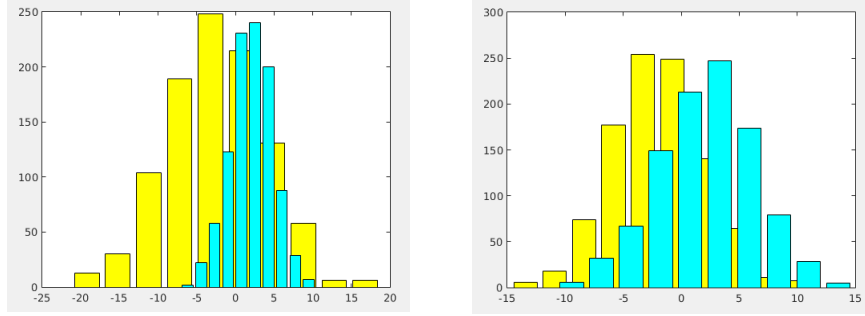
$$d_{ij} : \frac{P(X|W_i)}{P(X|W_j)} \begin{matrix} >_{w_i} \\ <_{w_j} \end{matrix} 1$$

Defining such classifiers, the feature function  $g_i(X)$  can be defined as :

$$\begin{aligned} g_1(X) &= \Pi_{j=2,3} \operatorname{sgn}\left(\frac{P(X|W_1)}{P(X|W_j)} - 1\right) \\ g_2(X) &= \Pi_{j=2,3} \operatorname{sgn}\left(\frac{P(X|W_2)}{P(X|W_j)} - 1\right) \\ g_3(X) &= \Pi_{j=2,3} \operatorname{sgn}\left(\frac{P(X|W_3)}{P(X|W_j)} - 1\right) \end{aligned}$$

In which the positive value of each function  $g_i(X)$  indicates that  $X$  belongs to  $W_i$  and vice versa.





(a) Projected distribution of data. (b) Projected distribution of data. The projected distribution of class 1 w.r.t. mapping vector  $V$  is displayed in blue color and the that of class 2 is displayed in yellow when  $s = 1$  is displayed in yellow when  $s = 0.5$  (BEST CASE). (WORST CASE).

Figure 5: A comparison between projected distributions over matrices  $V$  w.r.t. selected  $s$

- 
2. Express your discriminant functions in the form of linear discriminant functions.

The  $d_{ij}$ s could be expressed as linear functions as perpendicular bisectors of linking line between the center points of each class:

$$\begin{aligned} d_{12} &= X_2 + 3X_1 + 3 \\ d_{13} &= X_2 + \frac{1}{2}X_1 - \frac{3}{4} \\ d_{23} &= X_2 + 2X_1 - \frac{9}{2} \end{aligned}$$

Defining  $d_{ij}(X) = -d_{ji}(X)$ , the vector  $X$  belongs to class  $W_i$  iff  $\forall j \neq i, d_{ij}(X) > 0$

- 
3. Determine and plot the decision boundaries. Figure 6 displays the distribution of data given mean vectors and covariance matrices.

Figure 7 displays the data distribution boundary and derived boundaries using two-class discriminant functions defined in section 2 of this problem.

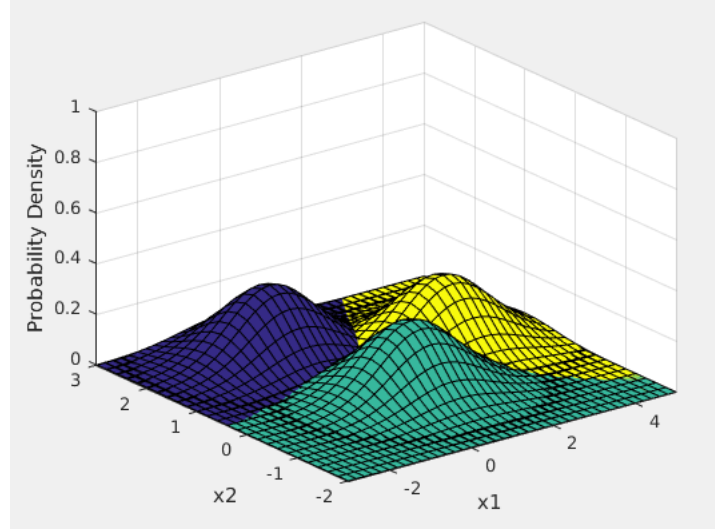


Figure 6: Data distribution w.r.t. means and covariance matrices

## 4 Problem 4

Consider the following 2-class classification problem involving a single feature  $x$ . Assume equal class priors and 0-1 loss function.

$$p(x|w_1) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}, p(x|w_2) = \begin{cases} 2 - 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

1. Sketch the two densities Since the class priors are equal:

$$p(w_1|x) = \frac{p(x|w_1)p(w_1)}{\sum_{i=1}^2 p(x|w_i)p(w_i)} = \frac{0.5(2x)}{1} = x$$

$$p(w_2|x) = \frac{p(x|w_2)p(w_2)}{\sum_{i=1}^2 p(x|w_i)p(w_i)} = \frac{0.5(2 - 2x)}{1} = 1 - x$$

2. State the Bayes decision rule and show the decision boundary.

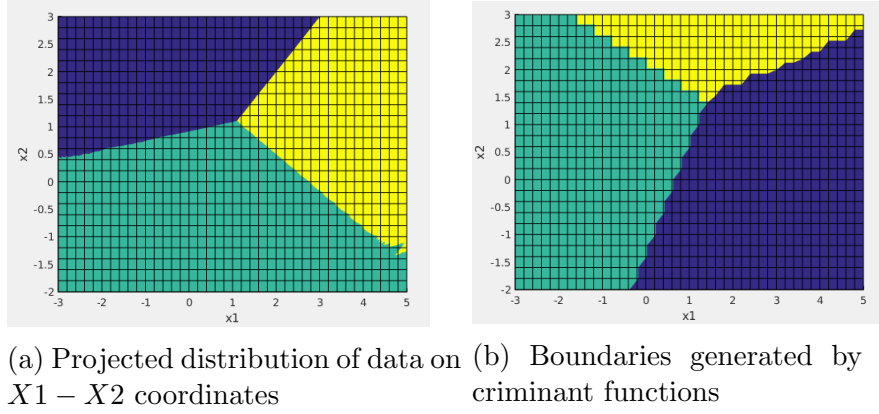


Figure 7: A comparison between projected distributions over matrices  $V$  w.r.t. selected  $s$

$$p(w_1|x) \underset{w_2}{\overset{w_1}{>}} p(w_2|x) \rightarrow x \underset{w_2}{\overset{w_1}{>}} 1 - x \rightarrow 2x - 1 \underset{w_2}{\overset{w_1}{>}} 0$$

Figure 8 displays the data distribution, decision boundary, false positive and false negative errors. The area coloured blue and green are displaying correct classifications for class  $w_1$  and  $w_2$  respectively. The yellow and red areas are illustrating  $\epsilon_1$  and  $\epsilon_2$  respectively.

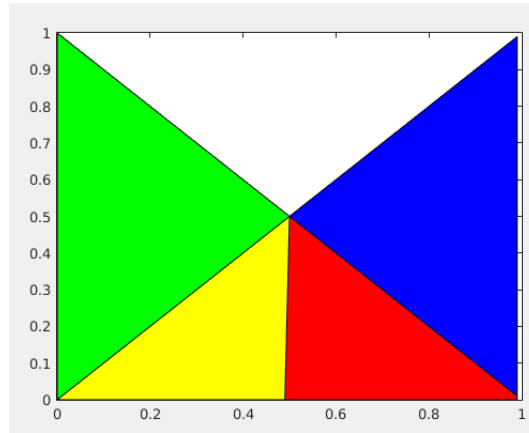


Figure 8: Bayes decision boundary for proposed data in problem 4 when  $p(w_1) = p(w_2) = 0.5$

3. What is the Bayes classification error? The classification error is being calculated as a weighted sum of each misclassified point w.r.t. its probability. Therefore,

$$\epsilon = 0.5 * \epsilon_1 + 0.5 * \epsilon_2 = 0.5 * \int_0^0 .5x dx + 0.5 * \int_0^1 .5(1-x) dx = \frac{1}{16} + \frac{1}{16} = \frac{1}{8}$$


---

4. How will the decision boundary change if the prior for class  $w_2$  is increased to 0.6?

$$p(w_1|x) = \frac{p(x|w_1)p(w_1)}{\sum_{i=1}^2 p(x|w_i)p(w_i)} = \frac{0.4(2x)}{1} = 0.8x$$

$$p(w_2|x) = \frac{p(x|w_2)p(w_2)}{\sum_{i=1}^2 p(x|w_i)p(w_i)} = \frac{0.6(2-2x)}{1} = 0.6 - 0.6x$$

$$p(w_1|x) \underset{w_2}{\overset{w_1}{>}} p(w_2|x) \rightarrow 0.4x \underset{w_2}{\overset{w_1}{>}} 0.6 - 0.6x \rightarrow x - 0.6 \underset{w_2}{\overset{w_1}{>}} 0$$

Figure 9 displays the data distribution, decision boundary, false positive and false negative errors. The area coloured blue and green are displaying correct classifications for class  $w_1$  and  $w_2$  respectively. The yellow and red areas are illustrating  $\epsilon_1$  and  $\epsilon_2$  respectively.

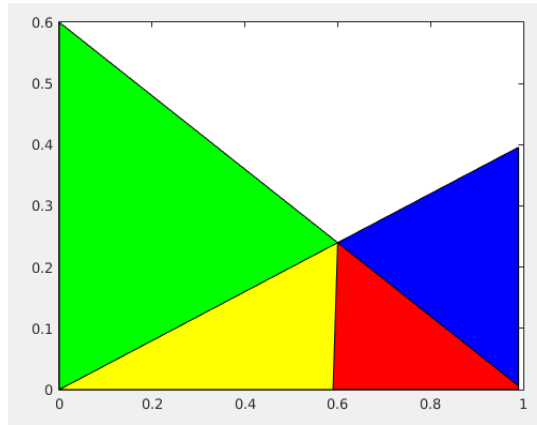


Figure 9: Bayes decision boundary for proposed data in problem 4 when  $p(w_1) = 0.4, p(w_2) = 0.6$

The classification error is being calculated as a weighted sum of each misclas-

sified point w.r.t. its probability. Therefore,

$$\epsilon = 0.4 * \epsilon_1 + 0.6 * \epsilon_2 = 0.4 * \int_0^1 .60.4x dx + 0.6 * \int_0^1 .6^1 0.6 - 0.6x dx = 0.0768$$


---

## 5 Problem 5

Consider a two-category classification problem in two dimensions with:

$$p(X|W_1) \propto N(0, I), P(X|W_2) \propto N\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, I\right), P(W_1) = P(W_2) = 0.5.$$

1. Calculate the Bayes decision boundary.

As  $\Sigma_1 = \Sigma_2 = \Sigma = I$ , a one dimensional function  $h(x)$  can be defined as bellow discriminating two classes linearly:

$$h(x) = (M_2 - M_1)^T \Sigma^{-1} X + \frac{1}{2} (M_1^T \Sigma^{-1} M_1 - M_2^T \Sigma^{-1} M_2) \underset{<_{w_2}}{\overset{>_{w_1}}{}} \ln \frac{P_1}{P_2}$$

So the appropriate function is as follows:

$$\begin{aligned} h(x) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T X + \frac{1}{2} \left( - \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) \underset{<_{w_2}}{\overset{>_{w_1}}{}} 0 \\ \rightarrow h(x) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T X - \frac{1}{2} \underset{<_{w_2}}{\overset{>_{w_1}}{}} 0 \end{aligned}$$


---

2. Calculate the Bhattacharyya error bound.

The Bhattacharyya error bound is in the form of following equation:

$$\epsilon = \sqrt{P(W_1)P(W_2)} e^{-\mu(0.5)} \quad (8)$$

Where  $\mu(0.5)$  for normal distribution is:

$$\mu(0.5) = \frac{1}{8} (M_2 - M_1)^T \left( \frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (M_2 - M_1) + \frac{1}{2} \ln \frac{\left| \frac{\Sigma_1 + \Sigma_2}{2} \right|}{\sqrt{|\Sigma_1| |\Sigma_2|}} \quad (9)$$

Considering  $\Sigma_1 = \Sigma_2 = I$  :

$$\mu(0.5) = \frac{1}{8}(M_2 - M_1)^T(M_2 - M_1) = \frac{1}{8}$$

So:

$$\epsilon = \sqrt{P(W_1)P(W_2)}e^{-\mu(0.5)} = \frac{1}{2}e^{-\frac{1}{8}}$$

3. Repeat the above for the same prior probabilities, but

$$p(X|W_1) \propto N(0, \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}), p(X|W_2) \propto N(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix})$$

As  $\Sigma_1 \neq \Sigma_2$ , the one dimensional discriminant function  $h(x)$  is formed as bellow:

$$\begin{aligned} h(x) &= \frac{1}{2}(x - M_1)^T \Sigma_1^{-1}(x - M_1) - \frac{1}{2}(x - M_2)^T \Sigma_2^{-1}(x - M_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \\ &\rightarrow h(x) = \frac{1}{3.75} x^T \begin{bmatrix} 2 & -0.5 \\ -0.5 & 2 \end{bmatrix} x - \frac{1}{21} (x - \begin{bmatrix} 1 \\ 0 \end{bmatrix})^T \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} (x - \begin{bmatrix} 1 \\ 0 \end{bmatrix}) \\ &\quad + \frac{1}{2} \ln \frac{|\begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}|}{|\begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}|} \\ &\rightarrow h(x) = \frac{1}{3.75} x^T \begin{bmatrix} 2 & -0.5 \\ -0.5 & 2 \end{bmatrix} x - \frac{1}{21} (x - \begin{bmatrix} 1 \\ 0 \end{bmatrix})^T \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} (x - \begin{bmatrix} 1 \\ 0 \end{bmatrix}) + \frac{3.75}{42} \\ &\rightarrow h(x) = \frac{1}{3.75} x^T \begin{bmatrix} 2 & -0.5 \\ -0.5 & 2 \end{bmatrix} x - \frac{1}{21} (x^T \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} x - x^T \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &\quad - \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} x - \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}) \\ &\rightarrow h(x) = \frac{1}{78.75} x^T \begin{bmatrix} 23.25 & -3 \\ -3 & 23.25 \end{bmatrix} x + \frac{1}{21} x^T \begin{bmatrix} 5 \\ -2 \end{bmatrix} + \frac{1}{21} \begin{bmatrix} 5 \\ -2 \end{bmatrix}^T x + \frac{1}{21} \begin{bmatrix} 29 \\ -20 \end{bmatrix}^T I + \frac{3.75}{42} \\ &\rightarrow h(x) = \frac{1}{78.75} x^T \begin{bmatrix} 23.25 & -3 \\ -3 & 23.25 \end{bmatrix} x + \frac{1}{21} \begin{bmatrix} 10 \\ -4 \end{bmatrix}^T x + \frac{21.75}{42} \end{aligned}$$

So the discriminant function is such as bellow:

$$h(x) = \frac{1}{78.75} x^T \begin{bmatrix} 23.25 & -3 \\ -3 & 23.25 \end{bmatrix} x + \frac{1}{21} \begin{bmatrix} 10 \\ -4 \end{bmatrix}^T x + \frac{21.75}{42} \begin{matrix} >_{w_1} 0 \\ <_{w_2} 0 \end{matrix}$$

The Bhattacharyya error bound is as bellow:

$$\begin{aligned}\mu(0.5) &= \frac{1}{8}(M_2 - M_1)^T \left( \frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (M_2 - M_1) + \frac{1}{2} \ln \frac{|\frac{\Sigma_1 + \Sigma_2}{2}|}{\sqrt{|\Sigma_1| |\Sigma_2|}} \\ \mu(0.5) &= \frac{1}{8} \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \left( \frac{\begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix} + \begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}}{2} \right)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{1}{2} \ln \frac{|\frac{\begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix} + \begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}}{2}|}{\sqrt{|\begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}| |\begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}|}} \\ \mu(0.5) &= \frac{3.5}{85.5} + 0.4948 = 0.5357\end{aligned}$$


---

## 6 Problem 6

Consider a two-category classification problem in two dimensions with  $p(W_1) = p(W_2)$ , assume that the class-conditional densities are Gaussian with mean  $\mu_1$  and co-variance  $\Sigma_1$  under class 1, and mean  $\mu_2$  and co-variance  $\Sigma_2$  under class 2. Further, assume that  $\mu_1 = \mu_2$ . (Note that in the problem set, there was stated that  $\mu_0 = \mu_1$ , which I guessed that the correct form is  $\mu_1 = \mu_2$  rather than  $\mu_0 = \mu_1$ !)

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

For the following case, draw contours of the level sets of the class conditional densities. Also, draw the decision boundaries obtained using the Bayes optimal classifier in each case and indicate the regions where the classifier will predict class 1 and where it will predict class 2.

Calculating the Bayes decision boundary yields:

$$h(X) = \frac{1}{2}(X - \mu_1)^T \Sigma_1^{-1} (X - \mu_1) - \frac{1}{2}(X - \mu_2)^T \Sigma_2^{-1} (X - \mu_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} \ln \frac{P(W_1)}{P(W_2)}$$

Considering the assumptions of the problem in addition to taking  $\mu_1 = \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$  yields:

$$\begin{aligned}
 h(X) &= \frac{1}{2}X^T \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}^{-1} X - \frac{1}{2}X^T \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}^{-1} X + \frac{1}{2} \ln \frac{\left| \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \right|}{\left| \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \right|} \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} 0 \\
 \rightarrow h(X) &= \frac{1}{8}X^T \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} X - \frac{1}{8}X^T \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} X \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} 0 \\
 \rightarrow h(X) &= \frac{1}{8} \begin{bmatrix} 1 \\ 4 \end{bmatrix} X^T X - \frac{1}{8} \begin{bmatrix} 4 \\ 1 \end{bmatrix} X^T X \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} 0 \\
 \rightarrow h(X) &= \frac{3}{8} \begin{bmatrix} -1 \\ 1 \end{bmatrix}^T (X^T X) \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} 0 \\
 \rightarrow h(X) &= \begin{bmatrix} -1 \\ 1 \end{bmatrix}^T (X^T X) \cdot \begin{matrix} >_{w_1} \\ <_{w_2} \end{matrix} 0
 \end{aligned}$$

A MATLAB script has been written to draw the required boundaries and regions. Figure 10 displays the result of this script.

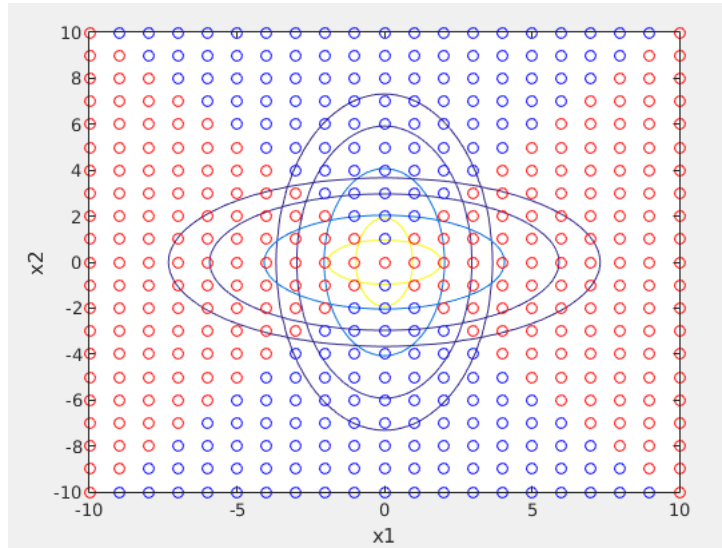


Figure 10: Bayesian decision boundaries and class acceptance regions for data described in problem 6



Contours of two classes are drawn in figure 10 and the acceptance region for each class has been displayed by coloured circles. The blue region is dedicated to the class  $W_1$  and the red region is dedicated to class  $W_2$ .

---

## 7 Problem 7

Consider the two-dimensional data points from two classes  $W_1$  and  $W_2$  below, and each of them come from a Gaussian distribution  $p(x|W_k) \propto N(\mu_k, \Sigma_k)$ . (Data table is not reported here)

1. What is the prior probability for each class, i.e.  $p(W_1)$  and  $p(W_2)$ .

Assuming that the given table represents the distribution of original data, the probability of each class is proportional to the ratio of observed data from that class in the given small dataset. So, the  $P(W_1) = \frac{6}{14} = 0.4286$  and therefore  $P(W_2) = 1 - P(W_1) = 0.5714$ .

---

2. Calculate the mean and covariance matrix for each class.

Estimating sample mean and sample covariance matrix from data is according to the following equations:

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_j \quad (10)$$

For the covariance matrix  $\Sigma^k = [\sigma_{ij}^k]$ , in which  $k$  denotes class number:

$$\hat{\sigma}_{ij}^k = \frac{1}{N_k - 1} \sum_{l=1}^{N_k} (x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j) \quad (11)$$

So according to these equations:

$$\hat{\mu}_1 = \begin{bmatrix} 1.5 \\ 1.3333 \end{bmatrix}, \hat{\mu}_2 = \begin{bmatrix} 8.25 \\ 8.625 \end{bmatrix} \quad (12)$$

and

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.1 & 1 \\ 1 & 1.8667 \end{bmatrix}, \hat{\Sigma}_2 = \begin{bmatrix} 2.7857 & 0.25 \\ 0.25 & 1.4107 \end{bmatrix} \quad (13)$$


---

3. Derive the equation for the decision boundary that separates these two classes, and plot the boundary. (Hint: you may want to use the posterior probability) Estimating class means and covariance matrices, the Bayesian decision boundary can be computed using the equations used in previous problems.

$$\frac{P(W_1|X)}{P(W_2|X)} \stackrel{>w_1}{<w_2} \frac{P(W_2)}{P(W_1)}$$

$$\frac{P(W_1|X)}{P(W_2|X)} = \frac{P(X|W_1).P(W_1)}{P(X|W_2).P(W_2)} = \frac{N\left(\begin{bmatrix} 1.5 \\ 1.3333 \end{bmatrix}, \begin{bmatrix} 1.1 & 1 \\ 1 & 1.8667 \end{bmatrix}\right).0.4286}{N\left(\begin{bmatrix} 8.25 \\ 8.625 \end{bmatrix}, \begin{bmatrix} 2.7857 & 0.25 \\ 0.25 & 1.4107 \end{bmatrix}\right).0.5714} \stackrel{>w_1}{<w_2} \frac{0.5714}{0.4286}$$

In which  $N(\mu, \Sigma)$  represents the Gaussian normal function for each class and with corresponding mean vector and covariance matrix.

4. Think of the case that the penalties for misclassification are different for the two classes (i.e. not zero-one loss), will it affect the decision boundary, and how?

Of course it will affect the decision boundary. Generally, the boundary is being tightened where the penalty of misclassification is less than other locations in favour of the locations in which the penalty of misclassification is larger. In other words, decision boundary will be adjusted in the way that make less misclassifications in the regions in which the penalty of misclassification is large, and prefer to make misclassifications in regions with less penalty.

---

## 8 Problem 8

Consider a classification problem with 2 classes and a single real-valued feature vector  $X$ . For class 1,  $p(x|c_1)$  is uniform  $U(a, b)$  with  $a = 2$  and  $b = 4$ . For class 2,  $p(x|c_2)$  is exponential with density  $\lambda \exp(-\lambda x)$  where  $\lambda = 1$ . Let  $p(c_1) = p(c_2) = 0.5$ .

1. Determine the location of the optimal decision regions

$$\begin{aligned}
 P(W_1|X)_{>_{w_2}^{w_1}} P(W_2|X) &\rightarrow P(X|W_1)P(W_1)_{>_{w_2}^{w_1}} P(X|W_2)P(W_2) \\
 &\rightarrow \frac{1}{4-2} \cdot_{>_{w_2}^{w_1}} e^{-x} \rightarrow -\ln(2) \cdot_{>_{w_2}^{w_1}} -x \\
 &\rightarrow x_{>_{w_2}^{w_1}} \ln(2)
 \end{aligned}$$

On the other hands, as  $p(X|W_1) \propto U(2, 4)$ , data from class 1 are not defined in  $R - [2, 4]$ . So there exists some implicit boundary conditions,  $\forall X \in W_1, X \in [2, 4]$ . Comparing the probability densities in  $X = \ln(2), X = 2, X = 4$  and everywhere  $X > 4$ , results choosing the region  $[2, 4]$  as the acceptance region for class 1 and every other region as the acceptance region for class 2.

This above-mentioned fact, is obviously clear in the figure 11, which represents the class densities and acceptance regions of classes.

2. Draw a sketch of the two class densities multiplied by  $P(c_1)$  and  $P(c_2)$  respectively, as a function of  $x$ , clearly showing the optimal decision boundary (or boundaries)

The figure 11, displays class densities multiplied by class priors along side optimal decision boundaries extracted from section 1 of this current problem.

The yellow area, is acceptance region for class 1. As it is clear, the class 1 is not defined in  $R - [2, 4]$  and in its domain, its density multiplied by its prior is greater than that of class 2. So the yellow region is an optimal boundary for this class. On the other hands, in  $R - [2, 4]$ , where class 1 is not defined, the optimal decision for data is obviously assigning that to class 2. The blue region in figure 11 displays optimal boundary for this class 2. The purple region displays the region of class 2 misclassifications.

3. Compute the Bayes error rate for this problem within 3 decimal places of accuracy

The Bayes error rate is computable as bellow:

$$\epsilon = \int_2^4 0.5 \exp(-x) dx = -0.5 \exp(-4) + 0.5 \exp(-2) = -0.5 * 0.018 + 0.5 * 0.135 = 0.058$$

4. Answer the questions above but now with  $a = 2$  and  $b = 22$ . Optimal decision

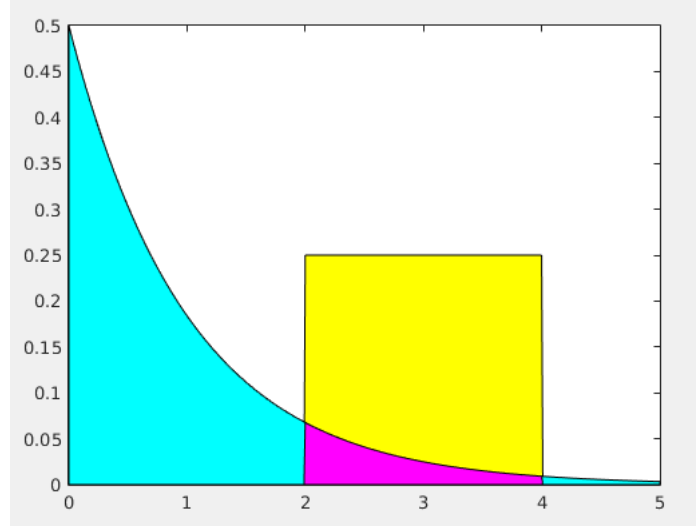


Figure 11: Class densities multiplied by class priors for each class given their conditional distributions in problem 8 when  $\text{dom } W_1 = [2, 4]$

boundary:

$$\begin{aligned}
 P(W_1|X)_{>_{w_2}^{w_1}} P(W_2|X) &\rightarrow P(X|W_1)P(W_1)_{>_{w_2}^{w_1}} P(X|W_2)P(W_2) \\
 &\rightarrow \frac{1}{22-2} \cdot_{<_{w_2}^{w_1}} e^{-x} \rightarrow -\ln(20) \cdot_{<_{w_2}^{w_1}} -x \\
 &\rightarrow x_{<_{w_2}^{w_1}} \ln(20) \rightarrow x_{<_{w_2}^{w_1}} 2.9957
 \end{aligned}$$

Since the  $x = 2.9957$  is in domain of class 1, the optimal point is itself and there is no need to compare it with boundary points of class 1.

Figure 12 displays the decision boundary for this properties. As it is clearly obvious, when  $2 \leq X \leq \ln(20)$ , the optimum decision for  $X$  is class 2 and when  $\ln(20) \leq X \leq 22$  the optimum decision is class 1 and after that when  $X > 22$ , as the class 1 is not defined there, the optimum decision is class 2 again.

Now, there is two types of error such that:

$$\epsilon = \epsilon_2^{\ln(20)} 0.5 \frac{1}{22-2} dx + \int_{\ln(20)}^2 20.5 \exp(-x) dx = \frac{1}{40} (\ln(20) - 2) + \frac{1}{2} (\exp(-22) - \frac{1}{20})$$

So, this is important to check the domain of each class in the cases in which the distribution of classes are finite and limited to a specific region.

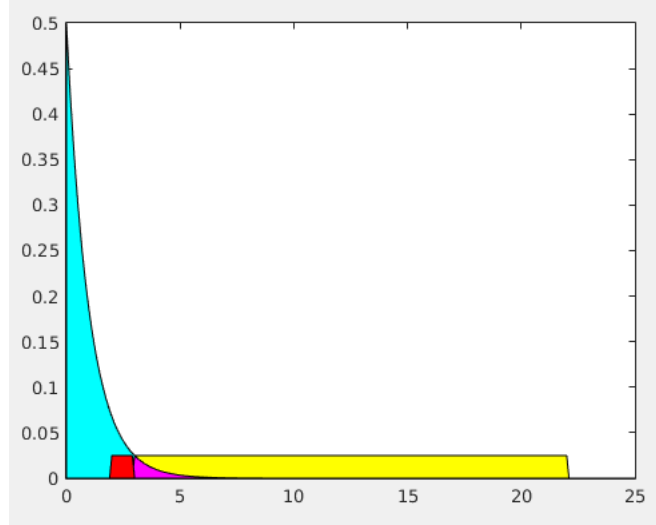


Figure 12: Class densities multiplied by class priors for each class given their conditional distributions in problem 8 when  $\text{dom } W_1 = [2, 22]$

## 9 Problem 9

Consider a 2-class classification problem with  $d$ -dimensional real-valued inputs  $x$ , where the class-conditional densities,  $p(x|c_1)$  and  $p(x|c_2)$  are multivariate Gaussian with different means  $\mu_1$  and  $\mu_2$  and a common covariance matrix  $\Sigma$ , with class probabilities  $P(c_1)$  and  $P(c_2)$ .

1. Write the discriminant functions for this problem in the form of  $g_1(x) = \log p(x|c_1) + \log p(c_1)$  (same for  $g_2(x)$ ).

The typical discriminant function could be specified as equation (14).

$$P(W_1|X) \underset{<_{w_2}}{>_{w_1}} P(W_2|X) \rightarrow \frac{P(X|W_1)P(W_1)}{P(X)} \underset{<_{w_2}}{>_{w_1}} \frac{P(X|W_2)P(W_2)}{P(X)} \quad (14)$$

So with the same  $P(X)$  for two classes, it will be in the form of equation (15)

$$P(X|W_1)P(W_1) \underset{<_{w_2}}{>_{w_1}} P(X|W_2)P(W_2) \quad (15)$$

Also, with respect to  $\log()$  function properties, it is possible to rewrite the equation (??) in the form of equation (16).

$$\begin{aligned} & \log(P(X|W_1)P(W_1)) \stackrel{w_1}{\underset{w_2}{>}} \log(P(X|W_2)P(W_2)) \\ \rightarrow & \log(P(X|W_1)) + \log(P(W_1)) \stackrel{w_1}{\underset{w_2}{>}} \log(P(X|W_2)) + \log(P(W_2)) \end{aligned} \quad (16)$$

So defining  $g_i(X)$  as follows yields good discriminant functions for each class.

$$g_i(X) = \log(P(X|W_i)) + \log(P(W_i)) \quad (17)$$

For the testing a new data, it is reasonable according to (16) and (17) to just compare the results of functions  $g_1(X)$  and  $g_2(X)$  and assign the  $X$  to the class which results larger value.

- 
2. Prove that the optimal decision boundary, at  $g(x) = g_1(x) - g_2(x) = 0$ , can be written in the form of a linear discriminant,  $w^T x + w_0 = 0$ , where  $w$  is a d-dimensional weight vector and  $w_0$  is a scalar, and clearly indicate what are  $w$  and  $w_0$  are in terms of parameters of the classification model.
- 

## 10 Problem 10

Consider the two-dimensional data points from two classes  $W_1$  and  $W_2$ . (The data table is not reported here)

1. Determine and plot the optimal projection line in a single dimension using Fisher linear discriminant method.

The sample mean vector and covariance matrix for each given data set has been calculated.

$$\mu_1 = \begin{bmatrix} 1.8334 \\ 1.8334 \end{bmatrix}, \mu_2 = \begin{bmatrix} 3.8334 \\ 3 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1.1 & 1 \\ 1 & 1.8667 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2.7857 & 0.25 \\ 0.25 & 1.4107 \end{bmatrix}$$

So:

$$V = (0.5\Sigma_1 + 0.5\Sigma_2)^{-1}(\mu_2 - \mu_1) = \frac{11.1725}{2} \begin{bmatrix} 3.8857 & 1.25 \\ 1.25 & 3.2774 \end{bmatrix} \begin{bmatrix} 2 \\ 1.1667 \end{bmatrix}$$

$$\rightarrow V = \frac{11.1725}{2} \begin{bmatrix} 9.223 \\ 6.3234 \end{bmatrix} = \begin{bmatrix} 51.5220 \\ 35.3241 \end{bmatrix}$$

And to compute optimum  $v_0$ , the iterative process should be accomplished. The script is implemented in MATLAB and the result is  $v_0 = 86.9$  with  $\epsilon = 0$ .

- 
2. Show the mapping of the points to the line as well as the Bayes discriminant assuming a suitable distribution.

Figure 13, displays required diagram.

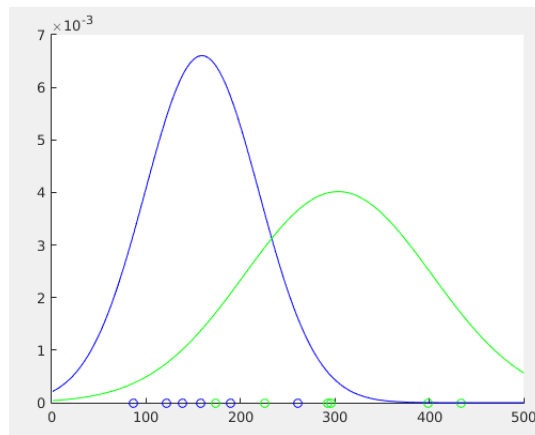


Figure 13: Mapped data points and the best fitting distributions for mapped data, problem 10.

---

## 11 Problem 11

[Computer Experiment] In this problem, you will be classifying the Iris dataset from the UCI Machine Learning Repository. The original data describes 3 classes of Iris flowers and contains various measurements about them:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

To make things a little easier, we left out one of the flower types and only included the Iris Setosa (class 0) and Iris Versicolor (class 1). There are two data files for this assignment, a training data set (70 data items) and a test data set (30 data items).

Every line in each file describes one data item; the 5 numbers in each row contain the 4 measurements in above order plus the class label. Use only the second measurement (sepal width) for classification and ignore all others.

---

The script of this problem is attached to the submitted homework zip file. In this report just the answers have been reported.

---

1. Load the training data. Compute a feature histogram for each class (see `histc` and `bar` in Matlab), using edges between the bins at 0.2 intervals between 1 and 5; i.e. `1 : 0.2 : 5`. Show the histograms either in two figures using subplot in Matlab, and make sure both plots have the same x-axis range.

Figure 14 displays the required subplot for feature histograms of both classes.

- 
2. Estimate the class priors from the training data. Report the prior probabilities of a flower being an Iris Setosa or being an Iris Versicolor.

$$P(W_0) = 0.5571 \text{ and } P(W_1) = 0.4429.$$

---

3. Assume that we can model the class-conditional probability density of each class using a univariate Gaussian (remember to only use the sepal width). What are the mean and variance of both class-conditional densities?



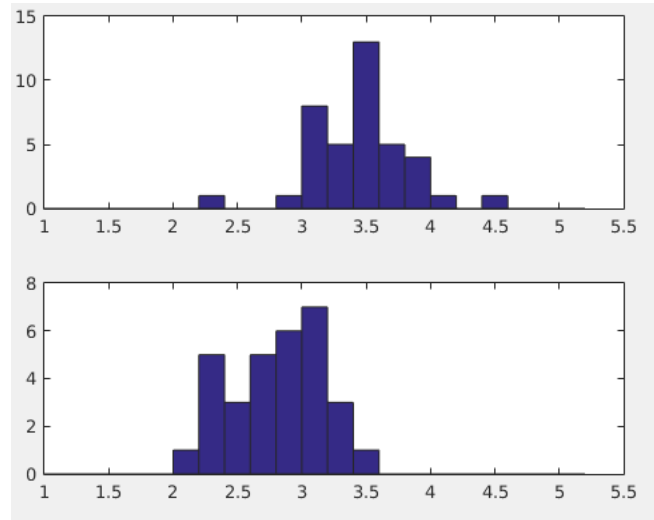
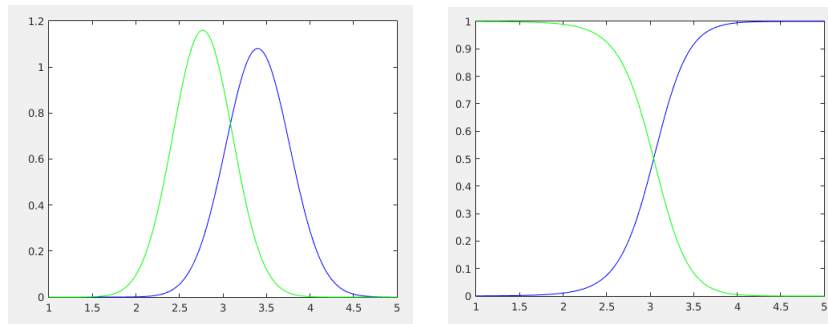


Figure 14: Histogram of features of two classes in train data set, problem 11.

- 
4. Plot the estimated class-conditional densities in a single diagram. Make a second figure plotting the class posteriors for both classes. Write down the equation you used to compute the class posteriors .  
Figure 15 illustrates the class-conditional densities.



(a) Estimated class-conditional densities (b) Estimated class posterior densities

Figure 15: Estimated class-conditional and class posterior densities, problem 11

To compute the class posterior the following equation has been used.

$$P(W_i|X) = \frac{P(X|W_i)P(W_i)}{\sum_{l=1}^2 P(X|W_l)P(W_l)} \quad (18)$$

- 
5. Compute the Bayes classification of the training data. How many Setosas do you mistakenly classify as Versicolor and how many Versicolors do you mistakenly classify as Setosa? What is the overall training error (i.e. percentage of incorrectly classified training data items)?

The error rate for class 0, which is the number of class 0 data misclassified as class 1 is 0.0513 and the error rate for class 1, which is the number of class 1 data misclassified as class 2 is 0.0645. The total error rate for training data is 0.0571.

- 
6. Now classify the test data. How many Setosas do you mistakenly classify as Versicolor and how many Versicolors do you mistakenly classify as Setosa? Report the number of cases that are incorrectly classified as above as well as the test error.

The error rate for class 0, which is the number of class 0 data misclassified as class 1 is 0.0256 and the error rate for class 1, which is the number of class 1 data misclassified as class 2 is 0.0645. The total error rate for testing data is 0.0428.

---