Shervin Halat

98131018

Statistical Pattern Recognition

Homework2

1.

First:

Notations:

BR = Bernie Sanders running for election

BR’ = Bernie Sanders not running for election

T = Trump winning // T ’ = Trump Losing

Assumptions:

P (T | BR) = 0.35 // P (T | BR’) = 0.65

a. P(BR) = 0.5 as it based on coin flip!

b. $P(T) = P(T \cap BR) + P(T \cap BR') =$

$= P(T \mid BR) * P(BR) + P(T \mid BR') * P(BR') = 0.5$

c. $P(BR \mid T') = P(BR \cap T') / P(T')$

i: $P(BR \cap T') = P(T' \mid BR) * P(BR) = (1 - P(T \mid BR)) * P(BR) =$

65/200 = 0.325

ii: P(T ‘) = 1 – 0.5 = 0.5

i & ii → P(BR | T ’) = 0.1625

Second:

Notations:

SBM: sleep before midnight

G: GDP more than 16

P(SBM) = 0.44 // P(G) = 0.52 // P(G ∩ SBM) = 0.37

d. what we want to show is in fact P(G|SBM):

P(G | SBM) = 0.37 / 0.44 = 0.84

The above result shows that the probability of getting good grades by getting enough sleep is around 0.84 which is a high chance of success.

Notations:

F: female // M: male // S: smoking

P(M) = 0.57 // P(S | M) = 14.6% // P(S | F) = 2.4%

e.

P(F) = 1 – 0.57 = 0.43

f.

P(S) = P(S|M) * P(M) + P(S|F) * P(F) = 0.0934

P(F | S) = P(S | F) * P(F) / P(S) = 2.4 / 100 * 0.43 / 0.0935 = 0.1103 = 11%

Third:

Notations:

PR: positive result // NR: negative result // PE: positive event // NE: negative event

false positive rate: 9.8% // false-negative rate: 1.4%
in other words: P(PR | NE) = 9.8% // P(NR | PE) = 1.4%

g.

P(PR | PE) = 1 – P(NR | PE) = 1 – 0.014 = 0.986

h.

assumptions: P(PE) = 0.02 // **question: P(PR) = ??**

The requested population fraction is the sum of both False Positives and True Positives results in proportion to mentioned population, which can be calculated by the following equation:

P(PR ∩ PE) + P(PR ∩ NE) = P(PR | PE) * P(PE) + P(PR | NE) * P(NE) =

0.986 * 0.02 + 0.098 * (1 – 0.02) = 0.11576 = **0.115 = P(PR)**

Hence, 0.115 fraction of the population will be considered as positive.

i.

question: P(PE | PR) = ??

P(PE | PR) = P(PR | PE) * P(PE) / P(PR) = 0.986 * 0.02 / 0.115 = 0.1714

j.

assumption: P(PR) = 0.5 // **question: P(PE | PR) = ??**

First we need to calculate P(PE ∩ PR):

P(PE ∩ PR) = P(PR | PE) * P(PE) = 0.986 * 0.02 = 0.0197

P(PE | PR) = P(PE ∩ PR) / P(PR) = 0.04

Therefore, probability that an athlete testing positive actually has used performance-enhancing drugs is **4 percent!!!**

k.

Assumption: P(PE) = prior probability

P(PE | PR) = posterior probability

Question: plot P(PE) vs P(PE | PR)

From previous parts we have:

P(PR = 0.115 & P(PR | PE) = 0.986

P(PE | PR) = P(PR | PE) * P(PE) / P(PR) = (0.986 / 0.115) * P(PE) → P(PE) = P(PE | PR) * 0.11

2.

a.

Features “B” and “b” are seem to be the best features as “b” separate silver from the two other classes and “B” separate bronze and gold.

b.

Covariance matrix= $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ **therefore:** $\sigma = 1$

$\mu_{Silver} = \begin{bmatrix} 101.2 \\ 1.5 \end{bmatrix}$ $\mu_{Bronze} = \begin{bmatrix} 105.8 \\ 41.8 \end{bmatrix}$ $\mu_{Gold} = \begin{bmatrix} 33 \\ 51.2 \end{bmatrix}$

➔ $G_{Silver(x)} = \frac{-1}{2}(x- \mu_{Silver})^T(x- \mu_{Silver})$

➔ $G_{Bronze(x)} = \frac{-1}{2}(x- \mu_{Bronze})^T(x- \mu_{Bronze})$

➔ $G_{Gold(x)} = \frac{-1}{2}(x- \mu_{Gold})^T(x- \mu_{Gold})$

Using the Bayes rule and considering the equal covariance matrixes and prior probabilities the discriminant function would be as follows:

Discriminant Function = Argmax $_{Class}$ **(G** $_{Class\ (x)}$**)**

for Class in (Silver, Bronze, Gold)

c.

in order to calculate the above equations for given data, python has been used. (**p2**)

```
Pixel 1:  G(silver)= -1930.94 G(bronze)= -380.74 G(gold)= -1516.32
Hence, for pixel 1 the appropriate Class is  bronze

Pixel 2:  G(silver)= -2096.64 G(bronze)= -1276.84 G(gold)= -7680.02
Hence, for pixel 2 the appropriate Class is  bronze

Pixel 3:  G(silver)= -3379.55 G(bronze)= -4519.84 G(gold)= -1259.62
Hence, for pixel 3 the appropriate Class is  gold

Pixel 4:  G(silver)= -475.44 G(bronze)= -1095.94 G(gold)= -6062.12
Hence, for pixel 4 the appropriate Class is  silver

Pixel 5:  G(silver)= -1896.45 G(bronze)= -985.14 G(gold)= -422.92
Hence, for pixel 5 the appropriate Class is  gold
```

According to the code results:

Pixel no.	**Class**
1	Bronze
2	Bronze
3	Gold
4	Silver
5	Gold

d.

The most important challenge of this system is that the number of training data for each class is less than the number of its features therefore considering less features seems a necessity. Another challenge is the assumption of Identity Matrix which ignores any covariances between two features and may not be acceptable. Besides, considering equal prior probability of classes seems too biased which directly affect outputs.

3.

Assumption: $P(W_0) = 0.4 \rightarrow P(W_1) = 0.6$

Knowing that: $\int_1^6 (P(x \mid w_0) = 1 \rightarrow P(x \mid w_0) = 0.2$

Knowing that: $\int_2^5 (P(x \mid w_2) = 1 \rightarrow P(x \mid w_0) = 0.33$

a.

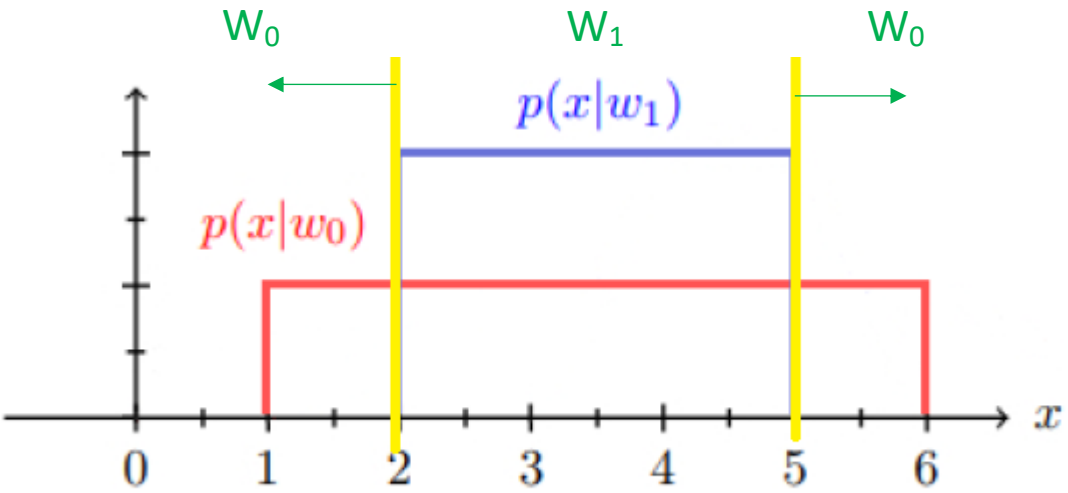


Figure 3 Class-conditional probability distributions

b.

Assuming symmetric decision boundaries and solving for one half of the plot we have:

$\xi_1 = 0.01 = P(\text{error} \mid w_1))$

$\xi_1 = 2 * \int_2^{2+d} (P(x \mid w_1)) = 0.01 \rightarrow 0.01 = 2 * 0.33 * d$

$\rightarrow d = 0.01515$

Total Error = $P(\text{error} \mid w_1) * P(W_1) + P(\text{error} \mid w_0) * P(W_0)$

$0.01 * 0.6 + P(w_0) * \int_{2+d}^{5-d} (P(x \mid w_0)) = 0.237$

→ **Total Error (**Neyman-Pearson**) = 0.243**

c.

d.

we have: $P(W_0) = 0.4$ & $P(W_1) = 0.6$

Bhattacharyya Error =

$\int_1^6 (P(x \mid w_0)^{1/2} * P(w_0)^{1/2} + P(x \mid w_1)^{1/2} * P(w_1)^{1/2})$

$= 5*(0.2^{0.5}*(0.4)^{0.5}) + 3*(0.33^{0.5}*(0.6)^{0.5}) = 2.75$

Bayes Error =

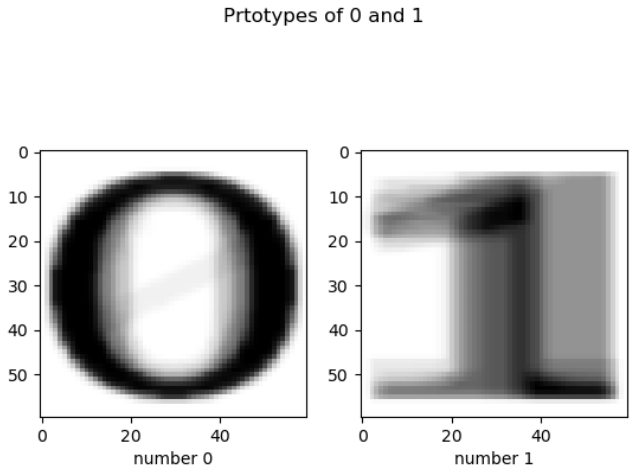
$\int_1^6 (\text{Min} \{ P(x \mid w_0) * P(w_0), P(x \mid w_1) * P(w_1) \}) =$

$\int_2^5 (P(x \mid w_0) * P(w_0)) = 3 * 0.4 * 0.2 = 0.24$

4.

a.

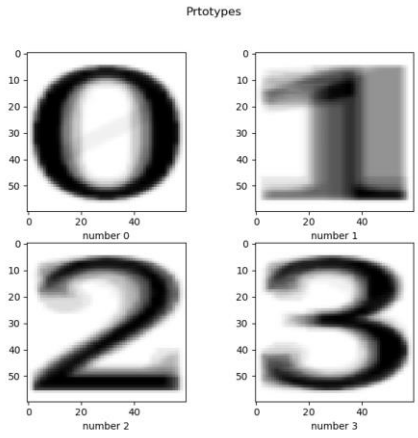
Prototypes of group “a”: (using code p4.a)



b. “Unable to compute Error due to Memory Error!”

c.

Prototypes of group “b”: (using code p4.c)

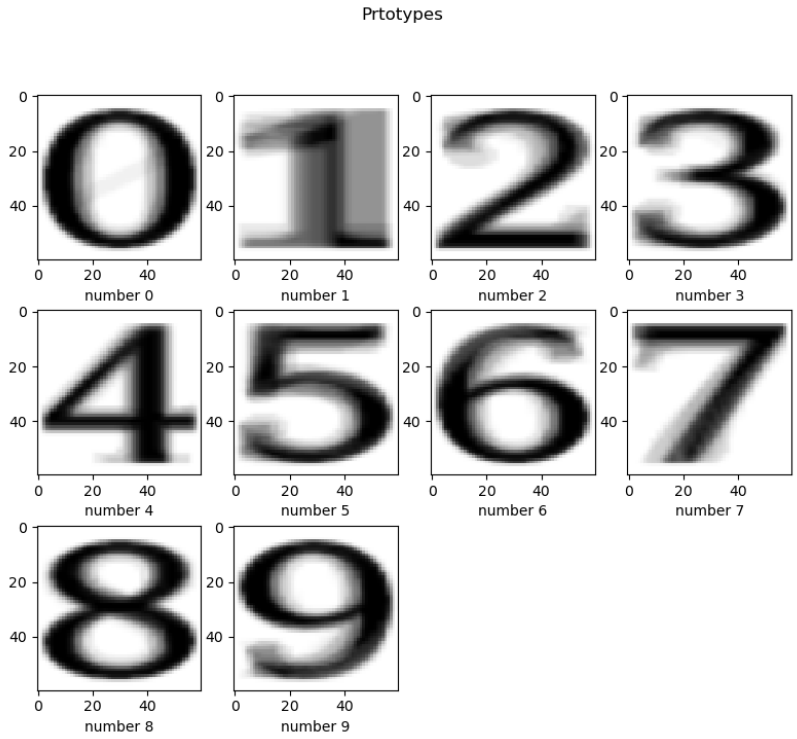


“Unable to compute Error due to Memory Error!”

d.

(Group “c” contains numbers from 0 to 10)

Prototypes of each class (numbers from 1 to 10) is shown in the figure below by code (p4.d):



“Unable to compute Error due to Memory Error!”

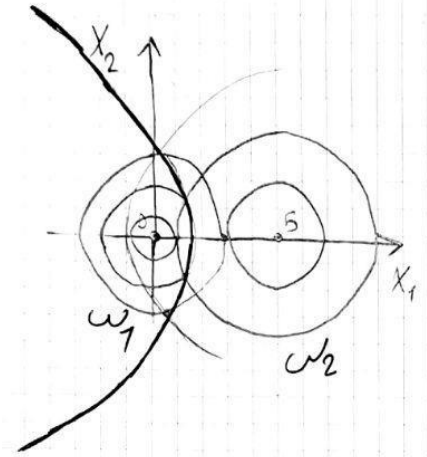
e.

5.

a.

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

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



b.

Assuming $P(\omega_1) = 3 \times P(\omega_2)$

$$g_1(x) = g_2(x) \Rightarrow \begin{cases} g_1(x) = -\frac{1}{2}\begin{bmatrix} X_1 & X_2 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} - \frac{1}{2} \times 2\ln(1) + \ln P(\omega_1) \\ g_2(x) = -\frac{1}{2} \times \frac{1}{4}\begin{bmatrix} X_1 - 5 & X_2 \end{bmatrix}\begin{pmatrix} X_1 - 5 \\ X_2 \end{pmatrix} - \frac{1}{2} \times 2\ln 4 + \ln P(\omega_2) \end{cases}$$

$g_1(x) = g_2(x) \Rightarrow$

Using Bayes Rule the Decision Boundry becomes:

$$3X_1^2 + 3X_2^2 + 10X_1 - 25 - 8\ln 12 = 0$$

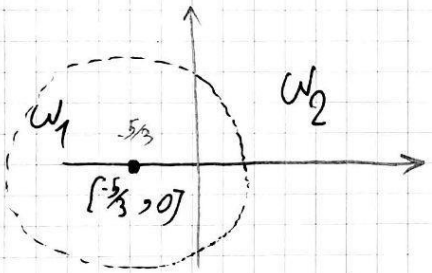
Therefore, the best Decision Boundry is quadratic as against $X_1 = 4$ (Linear)

c.

The Decision Boundry is quadratic (hyper-elipses)

From Previous Part we had:

$X_2^2 + (X_1 + \frac{5}{3})^2 = \alpha$ } The shape is a perfect Circle with radius of $\sqrt{\alpha}$ and Center of $[-\frac{5}{3}, 0]$



d.

Assuming $C_{12} = 5C_{21}$ and $C_{11} = C_{22} = 0$

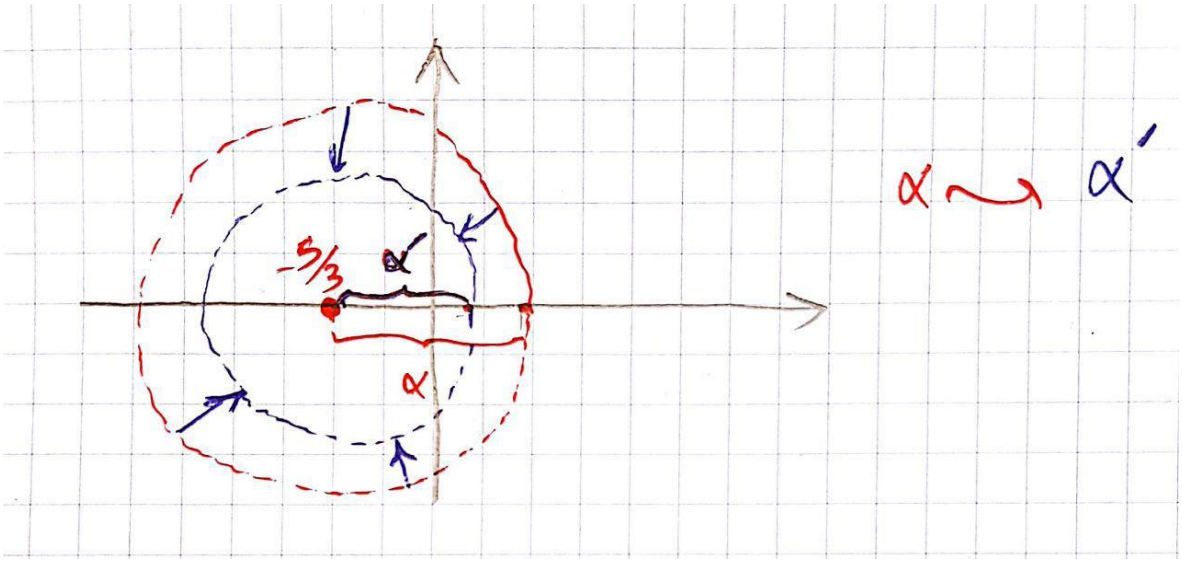
$\Longrightarrow$ Based on Bayes rule: $P(x|w_1)P(w_1)C_{21} \underset{w_2}{\overset{w_1}{\gtrless}} P(x|w_2)P(w_2)C_{12}$

Like Part "b" $\rightarrow$ $g_1'(x) = g_1(x) + \ln C_{21}$, $g_2'(x) = g_2(x) + \ln C_{12}$ $\xrightarrow{g_1'(x) = g_2'(x)}$ $\boxed{g_1(x) = g_2(x) + \ln 5}$

as we can see the only thing that has been changed by defining cost for misclassification is the radius of the circle of Part "c" which has been reduced.

α (from Part "c") $\rightarrow$ $\boxed{\alpha - \frac{2}{3}\ln 5 = \alpha'}$

Sketch of the change of discriminant function:



e.

f.

Calculation of Chernoff error bound needs parameter “s” which is not defined in the question!!

6.

a.

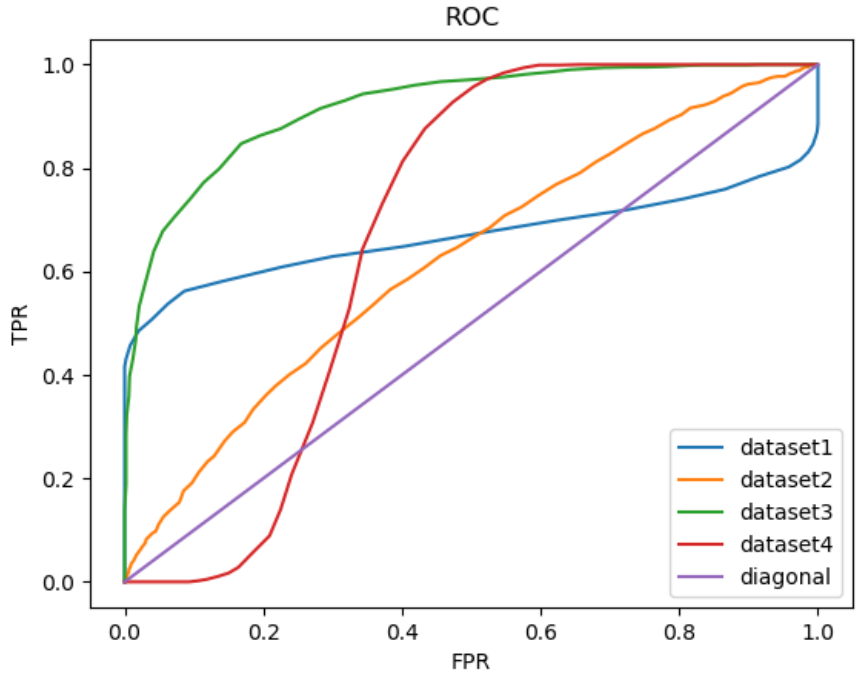
Using code (p6.a):

```
mean dataset 1 feature 1 = 9.997807
standard deviation dataset 1 feature 1 = 0.966031
mean dataset 1 feature 2 = 11.845612
standard deviation dataset 1 feature 2 = 4.033763
mean dataset 2 feature 1 = 10.225833
standard deviation dataset 2 feature 1 = 4.116033
mean dataset 2 feature 2 = 12.050214
standard deviation dataset 2 feature 2 = 4.133223
mean dataset 3 feature 1 = 10.039413
standard deviation dataset 3 feature 1 = 0.985593
mean dataset 3 feature 2 = 11.996726
standard deviation dataset 3 feature 2 = 0.998522
mean dataset 4 feature 1 = 10.149737
standard deviation dataset 4 feature 1 = 3.826697
mean dataset 4 feature 2 = 11.976080
standard deviation dataset 4 feature 2 = 0.983838
Discriminability of data set 1 = 0.445
Discriminability of data set 2 = 0.313
Discriminability of data set 3 = 1.395
Discriminability of data set 4 = 0.462
```

Dataset	μ_1	σ_1	μ_2	σ_2	Discriminability
1	9.99	0.96	11.84	4.03	0.445
2	10.22	4.11	12.05	4.13	0.312
3	10.03	0.98	11.99	0.99	1.395
4	10.14	3.82	11.97	0.98	0.462

b.

Using code (p6.b):

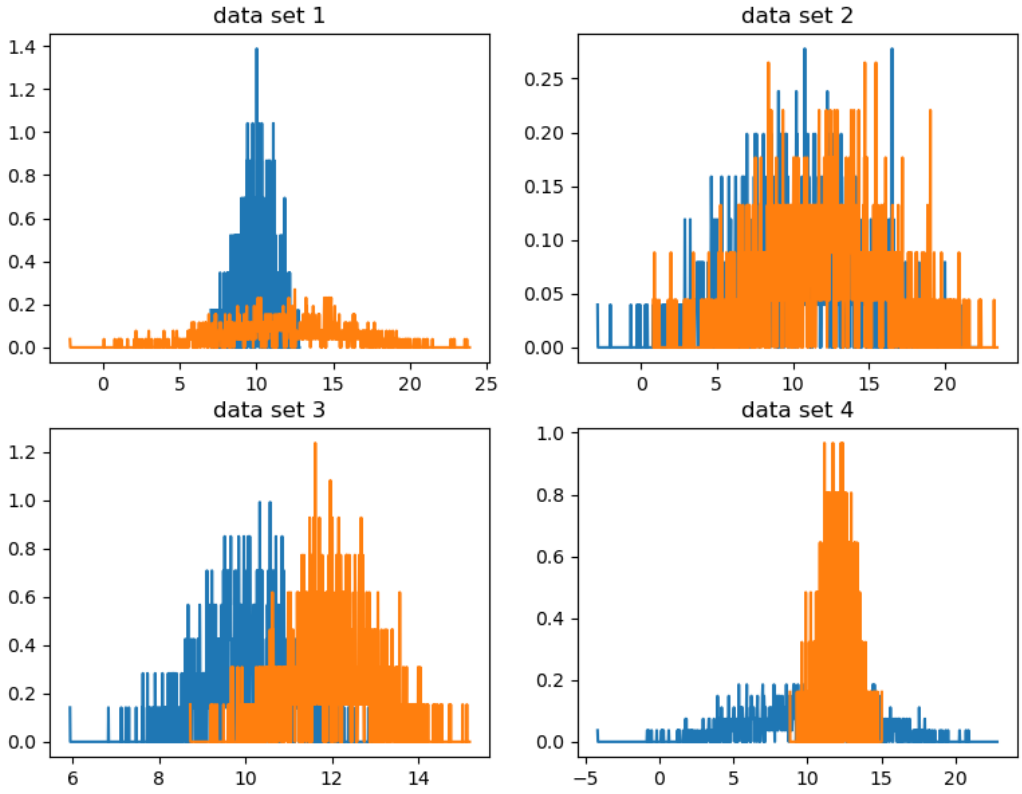


c.

Using the code (p6.c) the following plots were generated. For the purpose of generating PDF plots for each dataset. First, CDF of each Dataset were computed and then the PDFs were computed by an approximation of CDFs' derivate:

Pdf[i] = (cdf1[i]-cdf1[i+1])/(x1[i]-x1[i+1])

(The Monte Carlo method was implemented to find PDFs)



d.

According to the outputs of part “a”, Dataset 3 has the highest discriminability among all other datasets and as we knew before the better classifiers has the more AUC (Area Under Curve). By comparing datasets’ AUC of the plot in prat “b”, it’s obvious that dataset 3 has the maximum AUC which represent that this dataset has the best classifying ability among other datasets which is consistent with our observation of discriminability in part “a”. Another representation of mentioned conclusions are provided in part “c” as we can see that the two classes are more recognizable by comparison.

7.

a. P(face) = number of grey-level 0 pixels of masks / total number of pixels

   P(non-face) = 1 – P(face)

   face class prior = 0.221944

   non-face class prior = 0.778056

b.

8.

a.

Yes it is possible. This is actually how logistic regression, which is used for classification, is defined. The main equation for logistic regression, Sigmoid (Logistic) function, is generated based on Bayesian Theorem which is briefly mentioned in the following:

$$h_\theta(x) = g(\theta^T x)$$

$$z = \theta^T x$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

What sigmoid function(**g(z)**) calculates is actually the conditional probability of Y is 1 when feature vector **(X)** is given :

$$h_\theta(x) = P(y = 1|x; \theta) = 1 - P(y = 0|x; \theta)$$
$$P(y = 0|x; \theta) + P(y = 1|x; \theta) = 1$$

b.

Generally, Independence does not imply conditional independence and vice versa (also conditional independence does not imply independence). For example two famous counterexample for that are that two independent random

variables are rarely conditionally independent on their sum or on their maximum or anything like that.

For example, considering two independent random variables of X (any number from 1 to 10 equally likely) and Y (any number from 1 to 10 equally likely) and Assuming Z (Z = X + Y) as the condition. If X = 2 and Z = 5 then there is no chance for Y to be anything other than 3! Which shows that the value of Y is directly dependent to the value of X.

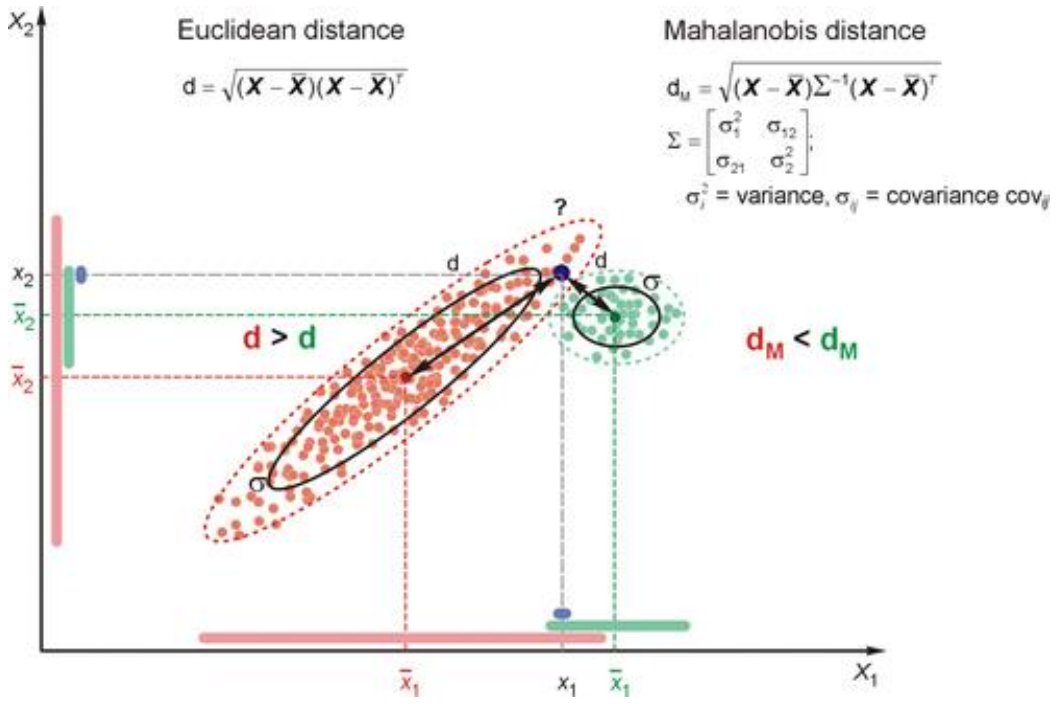
c.

Although Bayesian decision rule minimizes the error or in other words minimizes the probability of misclassification but in many cases that is not the point! That is, in many cases the cost of misclassifying one class against another may be remarkably much more. Therefore, the threshold may be shifted in order to lower the error corresponding to the class with higher cost of misclassification.
Another reason that No Free Launch Theorem also applies to Bayesian Decision Rule is that this method is based on availability of prior probabilities but in most cases a lot of information like priors are not available thus in such problems the Bayesian Decision Theory becomes worthless.

d.

Considering the two classes of the figure below, assuming we want to classify the black point with two MDC functions of Euclidean Distance or Mahalanobis Distance functions. As it is obvious the Black point is closer to the center of Green Class (cluster) which results in lower Euclidean Distance as opposed to Red Class (cluster). Therefore, classifying with Euclidean

Distance assign the Black Point to the Green Class which seems not logical. On the contrary, using Mahalanobis Distance function results in lower distance from center of Red Class as opposed to Green one. This is due to the fact that Mahalanobis Distance function takes covariance matrix of each class distribution into action or other words Mahalanobis Distance consider the distribution of all data points of each class when computing the distance, a point which Euclidean distance doesn't consider. Thus, using different MDC distance functions can lead to different outputs.



e.

Assuming that all prerequisites of Bayes Decision rule are available, such as mean, covariance and prior probabilities, the Bayes Decision rule's output would always be unique since the prerequisites which the Bayes' rule is based on are definite and unchangeable overtime.

f.

The training phase of Bayes' classifier consist computing distribution function (or variance for univariates) and prior probabilities of each class as against MDC classifier which its training phase is based on determining the Distance function and the parameters needed in that function (such as mean, variance or covariance) which can be varied.