IITM-CS5691 : Pattern Recognition and Machine Learning | Release Date: August 30, 2023
Assignment 1 | Due Date : September 14, 2023, 23:59

Roll No: CS23MO34 | Name: KUSHAGRA JAIN

1. (8 points) [GETTING YOUR BASICS RIGHT!]

   (a) (5 points) Let a random vector X follow a bivariate Gaussian distribution with mean $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance matrix $\Sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, i.e., $X \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a & b \\ c & d \end{bmatrix}\right)$. Then, use the pdf (probability density function) of X to:
   Find the distribution of (i) $X_2|X_1 = x_1$ and (ii) $X_1|X_2 = x_2$, and use them to (iii) find the permissible values of a, b, c, and d.
   (Hint: You can use the same approach of "completing the squares" seen in class).

   (b) (2 points) Consider the function $f(x) = x_1^2 + x_2^2 + x_1x_2$, and a point $v = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$. Find the linear approximation of f around v (i.e., $L_v[f](y)$), and show that the graph of this approximation is a hyperplane in $\mathbb{R}^3$.

   (c) (1 point) Which of these statements are true about two random variables X and Y defined on the same probability space?

      (i) If X, Y are independent, then X, Y are uncorrelated ($Cov(X, Y) = 0$).

      (ii) If X, Y are uncorrelated, then X, Y are independent.

      (iii) If X, Y are uncorrelated and follow a bivariate normal distribution, then X, Y are independent.

      (iv) None of the above.

Solution: (1) (a) (i)

The random variable x follows a bivariate Gaussian distribution with mean vector $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$. The covariance matrix $\boldsymbol{\Sigma}$ is symmetric, so $b = c$.

The inverse of $\boldsymbol{\Sigma}$ is given by $\boldsymbol{\Sigma}^{-1} = \frac{1}{ad-b^2} \begin{bmatrix} d & -b \\ -b & a \end{bmatrix}$.

The probability density function (PDF) of x is given by:

$$f_x(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Substituting the mean and covariance matrix values, we get:

$$f_x(x) = \frac{1}{(2\pi)\sqrt{ad-b^2}} \exp\left(\frac{-1}{2(ad-b^2)}\left[dx_1^2 + ax_2^2 - 2bx_1x_2\right]\right)$$

This can be further simplified to:

$$f_x(x) = \frac{1}{(2\pi)\sqrt{ad-b^2}} \exp\left(\frac{-1}{2(ad-b^2)}\left[\left(x_2 - \frac{b}{a}x_1\right)^2 + \frac{ad-b^2}{a^2}x_1^2\right]a\right)$$

Separating the terms, we have:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sqrt{d-\frac{b^2}{a}}} \exp\left(\frac{-1}{2\left(d-\frac{b^2}{a}\right)}\left[x_2 - \frac{b}{a}x_1\right]^2\right) \cdot \frac{1}{\sqrt{2\pi}\sqrt{a}} \exp\left(-\frac{1}{2a}x_1^2\right)$$

This implies that given $X_1 = x_1$, X_2 follows a bivariate Gaussian distribution with mean $\frac{b}{a}x_1$ and variance $d - \frac{b^2}{a}$. Additionally, X_1 follows a bivariate Gaussian distribution with mean 0 and variance a.

Solution: (1) (a) (ii)

The random variable x follows a bivariate Gaussian distribution with mean vector $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$. The covariance matrix $\mathbf{\Sigma}$ is symmetric, so $b = c$.

The inverse of $\mathbf{\Sigma}$ is given by $\mathbf{\Sigma}^{-1} = \frac{1}{ad-b^2} \begin{bmatrix} d & -b \\ -b & a \end{bmatrix}$.

The probability density function (PDF) of x is given by:

$$f_x(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Substituting the mean and covariance matrix values, we get:

$$f_x(x) = \frac{1}{(2\pi)\sqrt{ad-b^2}} \exp\left(\frac{-1}{2(ad-b^2)}\left[dx_1^2 + ax_2^2 - 2bx_1x_2\right]\right)$$

This can be further simplified to:

$$f_x(x) = \frac{1}{(2\pi)\sqrt{ad-b^2}} \exp\left(\frac{-1}{2(ad-b^2)}\left[\left(x_1 - \frac{b}{d}x_2\right)^2 + \frac{ad-b^2}{d^2}x_2^2\right] d\right)$$

Separating the terms, we have:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sqrt{a-\frac{b^2}{d}}} \exp\left(\frac{-1}{2\left(a-\frac{b^2}{d}\right)}\left[x_1 - \frac{b}{d}x_2\right]^2\right) \cdot \frac{1}{\sqrt{2\pi}\sqrt{d}} \exp\left(-\frac{1}{2d}x_2^2\right)$$

This implies that given $X_2 = x_2$, X_1 follows a bivariate Gaussian distribution with mean $\frac{b}{d}x_2$ and variance $a - \frac{b^2}{d}$. Additionally, X_2 follows a bivariate Gaussian distribution with mean 0 and variance d.

Solution: (1) (a) (iii)

To ensure that the given matrix $\mathbf{\Sigma} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$ represents a valid covariance matrix, we must satisfy the following conditions:

$$\begin{aligned} a &\geqslant 0 && \text{(Since variances are non-negative)} \\ b &= c && \text{(Due to symmetry in the covariance matrix)} \\ d &\geqslant 0 && \text{(Again, variances are non-negative)} \end{aligned}$$

In summary, the permissible values for a, b, c, and d are $a \geqslant 0$, $b = c$, and $d \geqslant 0$.

Solution: (1) (b)

We have the function $f(x) = x_1^2 + x_2^2 + x_1x_2$.

Given the point $v = \begin{bmatrix}3\\5\end{bmatrix}$, we can evaluate $f(v)$:

$$f(v) = 3^2 + 5^2 + 3 \times 5 = 49$$

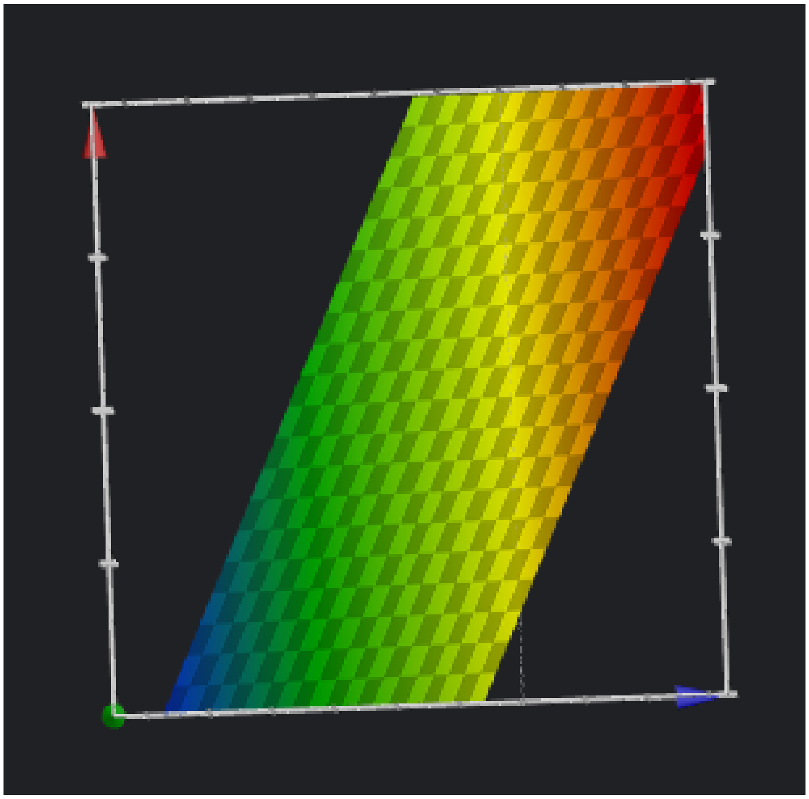
Taking partial derivatives with respect to x_1 and x_2:

$$\frac{\partial f(x)}{\partial x_1} = 2x_1 + x_2 \quad \text{and} \quad \frac{\partial f(x)}{\partial x_2} = 2x_2 + x_1$$

Now, let's consider $f(y) = f(x) + f'(x)\begin{bmatrix} y_1 - 3 \\ y_2 - 5 \end{bmatrix}$:

$$f(y) = 49 + [2x_1 + x_2 \quad 2x_2 + x_1]\begin{bmatrix} y_1 - 3 \\ y_2 - 5 \end{bmatrix} = 11y_1 + 13y_2 - 49$$

The graphical representation of this function is shown below:



From the graph, we observe that it forms a hyperplane in $\mathbb{R}^3$.

Solution: (c)
Only the first one is True.

2. (8 points) [EXPLORING MAXIMUM LIKELIHOOD ESTIMATION]
Consider the i.i.d data $\mathbf{X} = \{x_i\}_{i=1}^{n}$, such that each $x_i \sim \mathcal{N}(\mu, \sigma^2)$. We have seen ML estimates of μ, σ^2 in class by setting the gradient to zero.

   (a) (4 points) How can you argue that the stationary points so obtained are indeed global maxima of the likelihood function?

   (b) (4 points) Derive the bias of the MLE of μ, σ^2.

Solution: (a)
The Log Likelihood Function of a Univariate Gaussian Distribution is given by:

$$LL(\mu, \sigma / D_N) = -N \ln\left(\sqrt{2\pi\sigma}\right) - \sum_{i=1}^{n} \frac{(x^{(i)} - \mu)^2}{2\sigma^2}$$

Taking the derivative with respect to μ, we have:

$$\begin{aligned}
\frac{\partial LL}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(-N \ln\left(\sqrt{2\pi\sigma}\right) - \sum_{i=1}^{n} \frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right) \\
&= -\frac{N}{\sqrt{2\pi\sigma}} \cdot \frac{\partial \sigma}{\partial \mu} - \sum_{i=1}^{n} \frac{2(x^{(i)} - \mu)}{2\sigma^2} \cdot \frac{\partial \mu}{\partial \mu} \\
&= -\frac{N}{\sqrt{2\pi\sigma}} \cdot 0 - \sum_{i=1}^{n} \frac{2(x^{(i)} - \mu)}{2\sigma^2} \cdot 1 \\
&= \sum_{i=1}^{n} \frac{x^{(i)} - \mu}{\sigma^2}
\end{aligned}$$

Setting $\frac{\partial LL}{\partial \mu} = 0$, we get:

$$\sum_{i=1}^{n} \frac{x^{(i)} - \mu}{\sigma^2} = 0 \implies \boxed{\mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}}$$

Next, let's find the second derivative:

$$\frac{\partial^2 LL}{\partial \mu^2} = -\sum_{i=1}^{n} \frac{1}{\sigma^2}$$

Since $\frac{\partial^2 LL}{\partial \mu^2} < 0$, it indicates that it is a point of global maxima.

Solution: (b)
To derive the bias of the Maximum Likelihood Estimators (MLEs) of $\hat{\mu}$ and $\hat{\sigma}^2$, we'll need to consider the properties of MLEs.
The bias of an estimator $\hat{\theta}$ is defined as the expected value of the estimator minus the true parameter value:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

For the MLEs of $\hat{\mu}$ and $\hat{\sigma}^2$, denoted as $\hat{\mu}$ and $\hat{\sigma}^2$ respectively, we need to find:

$$Bias(\hat{\mu}) = E(\hat{\mu}) - \mu$$
$$Bias(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2$$

Starting with $\hat{\mu}$:

$$\begin{aligned}
E[\hat{\mu}] &= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} E[x_i] \\
&= \frac{1}{N}\sum_{i=1}^{N} \mu \\
&= \mu
\end{aligned}$$

Next, for $\hat{\sigma}^2$:

$$\begin{aligned}
E[\hat{\sigma}^2] &= E\left[\frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{N} (x_i - \hat{\mu})^2\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} x_i\hat{\mu} + \sum_{i=1}^{N} \hat{\mu}^2\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{N} x_i^2 - N\hat{\mu}^2 - N\hat{\mu}^2 + N\hat{\mu}^2\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{N} x_i^2 - N\hat{\mu}^2\right] \\
&= E[x^2] - E[\hat{\mu}^2]
\end{aligned}$$

$$
\begin{aligned}
&= \sigma_x^2 + E[x_n]^2 - \sigma_{\bar{\mu}}^2 - E[x_n]^2 \\
&= \sigma_x^2 - \sigma_{\bar{\mu}}^2 \\
&= \sigma_x^2 - \mathrm{Var}(\bar{\mu}) \\
&= \sigma_x^2 - \mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right) \\
&= \sigma_x^2 - \left(\frac{1}{N}\right)^2 \mathrm{Var}\left(\sum_{n=1}^{N} x_n\right) \\
&= \sigma_x^2 - \left(\frac{1}{N}\right)^2 N\sigma_x^2 \\
&= \sigma_x^2 - \frac{1}{N}\sigma_x^2 \\
&= \frac{N-1}{N}\sigma_x^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Bias}(\hat{\mu}) &= E(\hat{\mu}) - \mu \\
&= 0 \\
\mathrm{Bias}(\hat{\sigma}^2) &= E(\hat{\sigma}^2) - \sigma^2 \\
&= -\frac{1}{N}\sigma_x^2
\end{aligned}
$$

3. (8 points) [BAYESIAN DECISION THEORY]

(a) (4 points) [Optimal Classifier by Pen/Paper] Let L be the loss matrix defined by $L = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$, where L_{ij} indicates the loss for an input x with i being the true class and j the predicted class. Given the data:

x	-2.8	1.5	0.4	-0.3	-0.7	0.9	1.8	0.8	-2.4	-1.3	1.1	2.5	2.6	-3.3
y	1	3	2	2	1	3	3	2	1	1	2	3	3	1

find the optimal Bayes classifier $h(x)$, and provide its decision boundaries/regions. Assume that the class conditionals are Gaussian distributions with a known variance of 1 and unknown means (to be estimated from the data).

(b) (4 points) Consider a classification problem in which the loss incurred on mis-classifying an input vector from class C_k as C_j is given by loss matrix entry L_{kj}, and for which the loss incurred in selecting the reject option is ψ. Find the decision criterion that will give minimum expected loss, and then simplify it for the case of 0-1 loss (i.e., when $L_{kj} = \mathbb{1}_{k \neq j}$).

Solution: (a)
From the given data we can find that

Total points is 14
Class 1 point is 5
Class 2 point is 4
Class 3 point is 5

So, Prior Probability will be

$$p(Y=1) = \frac{5}{14}, \quad p(Y=2) = \frac{4}{14}, \quad p(Y=3) = \frac{5}{14}$$

$$\mu_1 = \frac{(-2.8 - 0.7 - 2.4 - 1.3 - 3.3)}{5} = -2.1$$

$$\mu_2 = \frac{(0.4 - 0.3 + 0.8 + 1.1)}{4} = 0.5$$

$$\mu_3 = \frac{(1.5 + 0.9 + 1.8 + 2.5 + 2.6)}{5} = 1.86$$

Variance for all classes is 1.

Posterior Probability will be

$$p(Y=1|\mathbf{X}) = \eta_1(\mathbf{X}) = \frac{p(\mathbf{X}|Y=1)\cdot p(Y=1)}{p(\mathbf{X}|Y=1)\cdot p(Y=1)+p(\mathbf{X}|Y=2)\cdot p(Y=2)+p(\mathbf{X}|Y=3)\cdot p(Y=3)}$$

$$\eta_1(\mathbf{X}) = \frac{\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\mathbf{X}-\mu_1)^2\right)\frac{5}{14}\right)}{\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\mathbf{X}-\mu_1)^2\right)\frac{5}{14}\right)+\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\mathbf{X}-\mu_2)^2\right)\frac{4}{14}\right)+\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\mathbf{X}-\mu_3)^2\right)\frac{5}{14}\right)}$$

$$\eta_1(X) = \frac{\exp\left(-\frac{1}{2}(X-\mu_1)^2\right)\cdot 5}{\exp\left(-\frac{1}{2}(X-\mu_1)^2\right)\cdot 5+\exp\left(-\frac{1}{2}(X-\mu_2)^2\right)\cdot 4+\exp\left(-\frac{1}{2}(X-\mu_3)^2\right)\cdot 5}$$

$$\eta_1(X) = \frac{5}{5+\exp\left(-\frac{1}{2}(X-\mu_2)^2+\frac{1}{2}(X-\mu_1)^2\right)4+\exp\left(-\frac{1}{2}(X-\mu_3)^2+\frac{1}{2}(X-\mu_1)^2\right)5}$$

$$\eta_1(X) = \frac{5}{5+\exp\left(-\left(X(\mu_1-\mu_2)+\frac{\mu_2^2-\mu_1^2}{2}\right)\right)4+\exp\left(-\left(X(\mu_1-\mu_3)+\frac{\mu_3^2-\mu_1^2}{2}\right)\right)5}$$

Now, putting the value of μ_1, μ_2 and μ_3

$$\eta_1(X) = \frac{5}{5+\exp\left(-\left(X(-2.1-0.5)+\frac{0.5^2-(-2.1)^2}{2}\right)\right)4+\exp\left(-\left(X(-2.1-1.86)+\frac{1.86^2-(-2.1)^2}{2}\right)\right)5}$$

$$\eta_1(X) = \frac{5}{5+\exp(2.6X+2.08)4+\exp(3.96X+0.4752)5}$$

Similarly,

$$\eta_2(X) = \frac{4}{4+\exp\left(-\left(X(\mu_2-\mu_1)+\frac{\mu_1^2-\mu_2^2}{2}\right)\right)5+\exp\left(-\left(X(\mu_2-\mu_3)+\frac{\mu_3^2-\mu_2^2}{2}\right)\right)5}$$

Now, putting the value of μ_1, μ_2 and μ_3

$$\eta_2(X) = \frac{4}{4+\exp\left(-\left(X(0.5-(-2.1))+\frac{(-2.1)^2-0.5^2}{2}\right)\right)5+\exp\left(-\left(X(0.5-1.86)+\frac{1.86^2-0.5^2}{2}\right)\right)5}$$

$$\eta_2(X) = \frac{4}{4+\exp(-2.6X-2.08)5+\exp(1.36X-1.6048)5}$$

Similarly,

$$\eta_3(X) = \frac{5}{5+\exp\left(-\left(X(\mu_3-\mu_1)+\frac{\mu_1^2-\mu_3^2}{2}\right)\right)5+\exp\left(-\left(X(\mu_3-\mu_2)+\frac{\mu_2^2-\mu_3^2}{2}\right)\right)4}$$

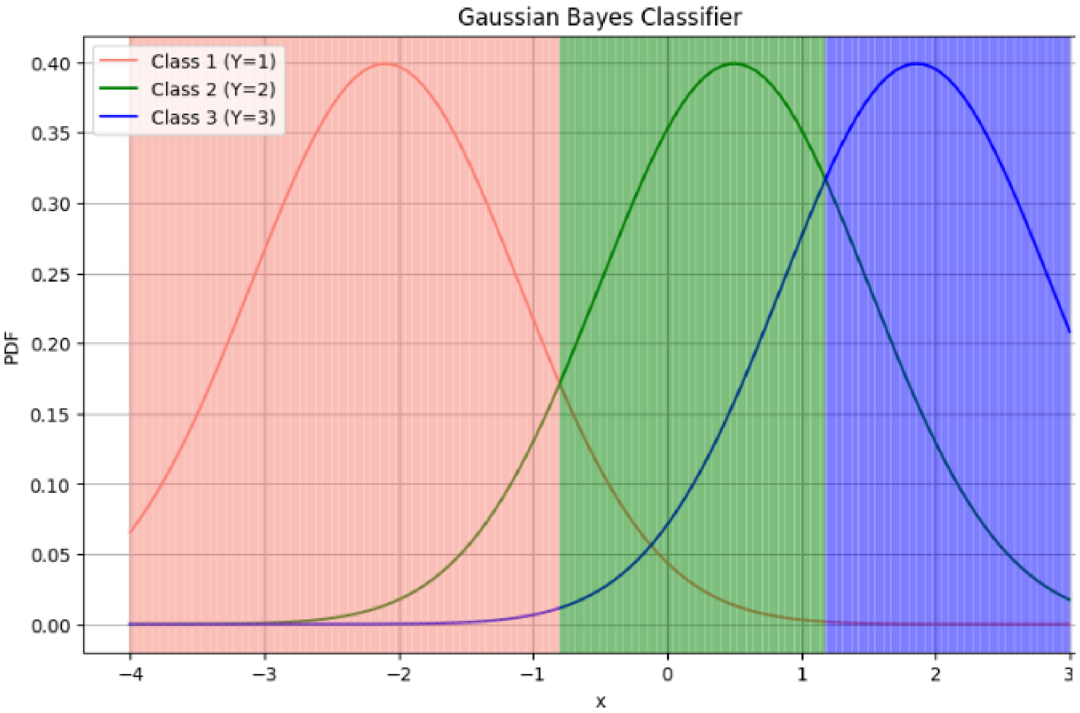
Now, putting the value of μ_1, μ_2 and μ_3

$$\eta_3(X) = \frac{5}{5+\exp\left(-\left(X(1.86-(-2.1))+\frac{(-2.1)^2-1.86^2}{2}\right)\right)5+\exp\left(-\left(X(1.86-0.5)+\frac{0.5^2-1.86^2}{2}\right)\right)4}$$

$$\eta_3(X) = \frac{5}{5+\exp(-3.6X-0.4752)5+\exp(-1.36X+1.6048)4}$$

Optimal Bayes Classifier h(x) :

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \eta_1(x) \\ \eta_2(X) \\ \eta_3(X) \end{bmatrix} = argmax \begin{bmatrix} \eta_2(X)+2\eta_3(X) \\ \eta_1(x)+\eta_3(X) \\ 2\eta_1(x)+\eta_2(X) \end{bmatrix}$$



Solution: (b)

Let $p(C_k|x)$ be the probability that a vector x belongs to class C_k. If we assign x to class C_j, then we will incur a loss of L_{kj}.

If we decide to classify x in C_k, the expected loss incurred by us will be given by:

$$\sum_k L_{kj} p(C_k|x)$$

This summation represents the losses incurred when x belongs to class k but is classified as C_j, class j but is classified as C_k, and so on.

Therefore, in order to minimize the expected loss, we choose to classify x into the class that has the minimum loss. Let's denote this class as j:

$$j = \arg\min_l \sum_k L_{kl} p(C_k \,|\, x)$$

Here, j is the class that results in the minimum loss. However, we also have the option to reject, in which case we incur a loss of ψ.

If $\psi \leqslant \lambda$, we will choose to classify x as the reject option.

$$\text{choose} \begin{cases} \text{class } j, & \text{if } \min l \sum_k Lklp\left(C_k \mid x\right) < \lambda \\ \text{reject}, & \text{otherwise.} \end{cases}$$

Now, if L is a 0-1 loss function, meaning $L_{ij} = 0$ if $i \neq j$, then $\sum_k L_{kj} p(C_k|x)$ simplifies to $1 - p(C_k|x)$.

So,

$$j = \arg\min_l (1 - p(C_k \,|\, x))$$

similarly like earlier

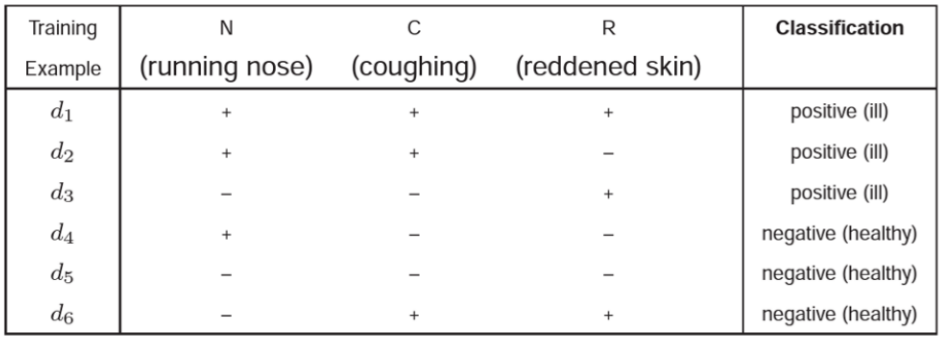
$$c \qquad\qquad \text{hoose} \begin{cases} \text{class } j, & \text{if } \arg\min_l 1 - p\left(C_k \mid x\right) < \lambda \\ \text{reject}, & \text{otherwise.} \end{cases}$$

4. (8 points) [REVEREND BAYES DECIDES FURTHER!]

(a) (2 points) For a two-class optimal Bayes classifer h, the decision region is given by: $R_i = \{x \in \mathbb{R} : h(x) = C_i\}$. Is R_1 always a single interval (based on a single cutoff separating the C_1 and C_2 class) or can R_1 be composed of more than one discontiguous interval? If yes for latter, give an example by plotting the pdfs $p(x, C_1)$ and $p(x, C_2)$ against x.

(b) (2 points) For a binary classifer h, let $L = \begin{bmatrix} p & q \\ r & s \end{bmatrix}$ be the loss matrix; and $C_{\text{train}} = \begin{bmatrix} 100 & 10 \\ 20 & 120 \end{bmatrix}$, and $C_{\text{test}} = \begin{bmatrix} 90 & 45 \\ 30 & 85 \end{bmatrix}$ be the confusion matrix when h is applied on the training and test data respectively. All three matrices have ground-truth classes t along the rows and predictions h along the columns in the same order for the two classes. Express your estimate of the expected loss of h in terms of p to s above.

(c) (4 points) Consider the dataset introduced in the table below, where the task is to predict whether a person is ill. We use a representation based on three features per subject to describe an individual person. These features are "running nose (N)", "coughing (C)", and "reddened skin (R)", each of which can take the value true ('+') or false ('–'). (i) Classify the data point ($d_7 : N = -, C = +, R = -$) using a Naive Bayes classifier. As part of your solution, also write down the (ii) Naive Bayes assumption and (iii) Naive Bayes classifier, along with (iv) which distribution's MLE formula you used to estimate the class conditionals.



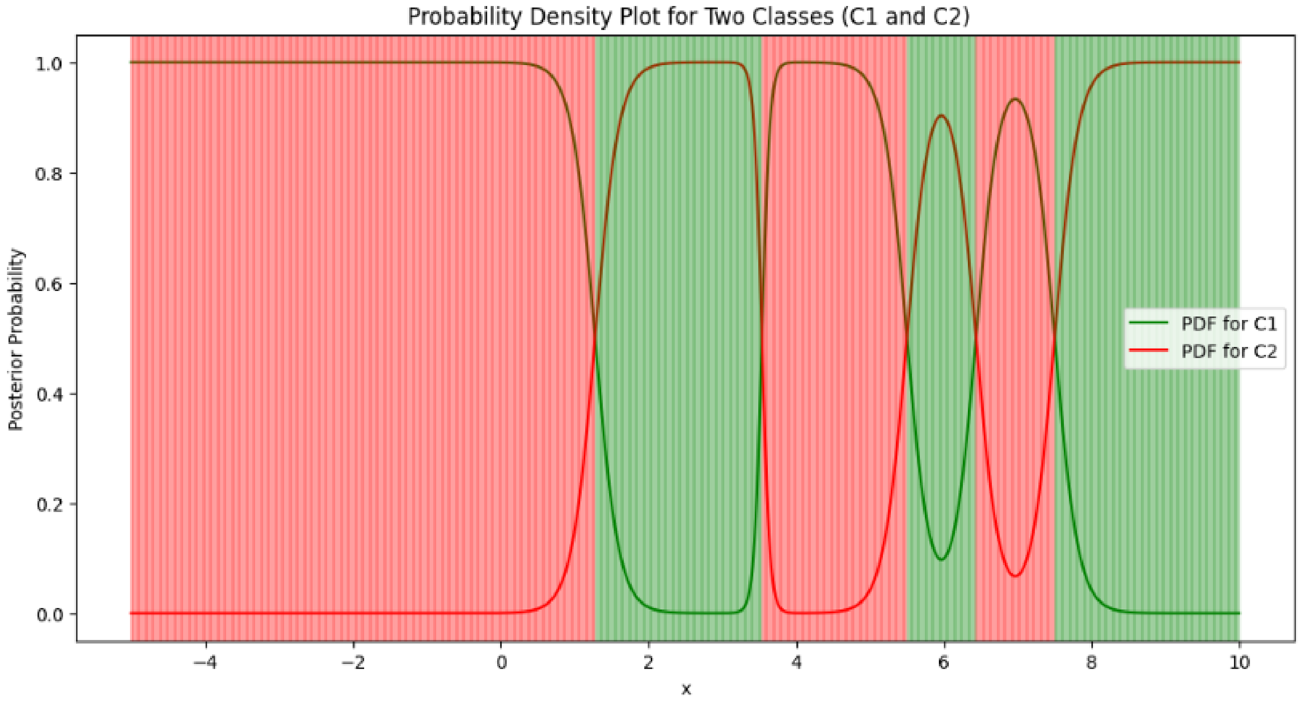
Training Example	N (running nose)	C (coughing)	R (reddened skin)	**Classification**
d_1	+	+	+	positive (ill)
d_2	+	+	–	positive (ill)
d_3	–	–	+	positive (ill)
d_4	+	–	–	negative (healthy)
d_5	–	–	–	negative (healthy)
d_6	–	+	+	negative (healthy)

Solution: (a)

The decision region R_1 for a two-class optimal Bayes classifier h may not always be a single interval. It can be composed of more than one discontiguous interval. This can happen when the probability density functions $p(x, C_1)$ and $p(x, C_2)$ overlap in a way that there are multiple disconnected regions where C_1 is more probable than C_2.

As an example, consider a scenario where $p(x, C_1)$ and $p(x, C_2)$ are two normal distributions with different means but similar variances. In this case, the decision boundary could be two disconnected intervals.

This condition ensures that the overlap between the two normal distributions is significant enough to create multiple disconnected regions where C_1 is more probable than C_2.



Solution: (b)
The Loss Matrix is Given by :

$L = \begin{bmatrix} p & q \\ r & s \end{bmatrix}$

The expected loss of a binary classifier h, given the loss matrix L can be calculated with the help of formula below:

$$E(L) = \sum_{i=1}^{N} \sum_{j=1}^{N} L_{ij} \cdot P(t_i, h_j)$$

here,

- $E(L)$ is the expected loss.
- L_{ij} is the element in the loss matrix L corresponding to the true class t_i and the predicted class h_j.
- $P(t_i, h_j)$ is the joint probability of true class t_i predicted class h_j.

Confusion matrix of Training data and Test Data is given as :

$$C_{\text{train}} = \begin{bmatrix} 100 & 10 \\ 20 & 120 \end{bmatrix} \qquad C_{\text{test}} = \begin{bmatrix} 90 & 45 \\ 30 & 85 \end{bmatrix}$$

To estimate the expected loss of classifier h in terms of the elements of the loss matrix L, we use the following formula:

$$E(L) = \frac{1}{N} \sum_{n=1}^{N} L_{t,h_x}$$

So,

$$\boxed{E(L) = \frac{1}{N}(p \cdot C_{\text{train}(1,1)} + q \cdot C_{\text{train}(1,2)} + r \cdot C_{\text{train}(2,1)} + s \cdot C_{\text{train}(2,2)})}$$

Given the confusion matrix C_{train} and the loss matrix L, we can substitute the values and calculate the expected loss.

$$E(L) = \frac{1}{250}(p \cdot 100 + q \cdot 10 + r \cdot 20 + s \cdot 120)$$

For C_{test}, we can follow the same procedure using the corresponding values. So, the expected loss for the test data would then be:

$$E(L) = \frac{1}{250}(p \cdot 90 + q \cdot 45 + r \cdot 30 + s \cdot 85)$$

This gives an estimate of the overall loss incurred by using classifier h on the test data.

Solution: (c)
The Following data contain two Class **ill** and **healthy**.

$p(\text{ill}) = \frac{1}{2} = 0.5$

$p(\text{healthy}) = \frac{1}{2} = 0.5$

The Prior probability of the data is given by :

.

Running Nose	Yes	No
+	2/3	1/3
-	1/3	2/3

Coughing	Yes	No
+	2/3	1/3
-	1/3	2/3

Reddened Skin	Yes	No
+	2/3	1/3
-	1/3	2/3

Given the data point d_7 with attributes $N = -$, $C = +$, and $R = -$, we want to classify it. We'll use the Naive Bayes classifier.

$$\begin{aligned}
p(\text{ill}/d_7) &= p(d_7/\text{ill}) \times p(\text{ill}) \\
&= p(N/\text{ill}) \times p(C/\text{ill}) \times p(R/\text{ill}) \times p(\text{ill}) \\
&= \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{2} \\
&= \frac{1}{27}
\end{aligned}$$

$$\begin{aligned}
p(\text{healthy}/d_7) &= p(d_7/\text{healthy}) \times p(\text{healthy}) \\
&= p(N/\text{healthy}) \times p(C/\text{healthy}) \times p(R/\text{healthy}) \times p(\text{healthy}) \\
&= \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{2} \\
&= \frac{2}{27}
\end{aligned}$$

Optimal Naive bayes classifier : $h(x) = argmax(1/27, 2/27)$

hence, $h(x)$= healthy.
So, this data point will be classified as **healthy**.

Here the Class Condition is estimated using **Bernaulli** distribution

.