IITM-CS5691 : Pattern Recognition and Machine Learning — Release Date: August 26, 2025
Assignment 1 — Due Date : September 17, 2025, 9pm

Roll No: EE25S009 — Name: RITABRATA MANDAL
Collaborators (if any):
References/sources (if any): Joram Soch et al. Statproofbook/statproofbook.github.io: The book of statistical proofs (version 2023). Zenodo, 2024.

General Instructions:

- Use LaTeX to write-up your solutions (in the solution blocks of the source LaTeX file of this assignment), and submit the resulting pdf files (one per question) at Gradescope by the due date. (Note: **No late submissions** will be allowed, other than one-day late submission with 10% penalty or four-day late submission with 30% penalty! You can join Gradescope using the course entry code 6K4P43 and submit your solution to each question within Gradescope as per instructions that will be emailed later).
- For the programming question, please submit your code (rollno.ipynb file and rollno.py file in rollno.zip) directly in moodle, but provide your results/answers (including Jupyter notebook **with output**) in the pdf file you upload to Gradescope.
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. Same rules apply for codes written for any programming assignments (i.e., write your own code; we will run plagiarism and AI detection checks on codes).
- If you have referred a book or any other online material for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words.
- For all the reasons explained in class, you cannot feed these questions into LLMs (Large Language Models like ChatGPT) and cannot use the LLMs' outputs to answer this assignment. Related to this, please also complete the self-declaration statement in the end of your answer sheet pdf.
- Please be advised that *the lesser your reliance on online materials or LLMs* for answering the questions, *the more your understanding* of the concepts will be and *the more prepared you will be for the course exams*.
- Points will be awarded based on how clear, concise and rigorous your solutions are, and how correct your answer is. The weightage of this assignment is 11% towards the overall course grade.

1. (8 points) [RANDOM(NESS RELATED) QUESTIONS]

   (a) (2 points) Clark Kent has lost his dog Krypto in either forest A (with a priori probability 0.4) or in forest B (with a priori probability 0.6). On any given day, if Krypto is in A and Clark spends a day searching for it in A, the conditional probability that he will find Krypto that day is 0.25. Similarly, if Krypto is in B and Clark spends a day looking for it there, the conditional probability that he will find Krypto that day is 0.15. Krypto cannot go from one forest to the

other. Clark can search only in the daytime, and he can travel from one forest to the other only at night.

i. (1 point) In which forest should Clark look to maximize the probability he finds Krypto on the first day of the search?

**Solution:** Define the events
A = event that Clark has lost Krypto in forest A
B = event that Clark has lost Krypto in forest B
F = event that he found Krypto
We know, $p(A) = 0.4, p(B) = 0.6, p(F \mid A) = 0.25, p(F \mid B) = 0.15$ we have to find out the maximum value among $p(A \mid F)$ and $p(B \mid F)$ To do that by Bayes rule we have

$$p(A \mid F) = \frac{p(A)p(F \mid A)}{p(F)}; \quad p(B \mid F) = \frac{p(B)p(F \mid B)}{p(F)}$$
$$\Rightarrow p(A \mid F) = 0.1/p(F); \qquad p(B \mid F) = 0.09/p(F)$$

As $p(A \mid F) > p(B \mid F)$ we conclude that it it better to search in forest A.

ii. (1 point) Given that Clark looked in A on the first day but didn't find Krypto, what is the probability that Krypto is in A?

**Solution:** Define events
A = Krypto is in forest A.
B = Krypto is in forest B.
F = Clark searches in forest A on the first day and does not find Krypto.
We are asked to compute $P(A \mid F)$.

$$P(A \mid F) = \frac{P(F \mid A)\, P(A)}{P(F \mid A)\, P(A) + P(F \mid B)\, P(B)}.$$

And we can write

$$P(F \mid A) = 1 - 0.25 = 0.75,$$

since if Krypto is in forest A, the probability Clark does not find him in one day is 0.75. And

$$P(F \mid B) = 1,$$

since if Krypto is in forest B, Clark is searching the wrong forest and will certainly not find him. Also given

$$P(A) = 0.4, \quad P(B) = 0.6.$$

Finally,

$$P(A \mid F) = \frac{0.75 \times 0.4}{0.75 \times 0.4 + 1 \times 0.6} = \frac{1}{3}.$$

(b) (2 points) Consider the trivariate Gaussian distribution,

$$p(x_1, x_2, x_3) = \mathcal{N}\left( \begin{bmatrix} 0 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 & +1 \\ -1 & 5 & -1 \\ +1 & -1 & 10 \end{bmatrix} \right).$$

Compute the following:

i. (1 point) $p(x_1, x_2 | x_3 = 1)$.

**Solution:** Let, $X = \begin{bmatrix} X_a \\ X_b \end{bmatrix}$, where $X_a = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $X_b = [x_3]$
This leads to following partitions

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} = \left[\begin{array}{c} 0 \\ 2 \\ \hline 4 \end{array}\right] \text{ and } \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} = \left[\begin{array}{cc|c} 0.3 & -1 & 1 \\ -1 & 5 & -1 \\ \hline 1 & -1 & 10 \end{array}\right]$$

We know that, the conditional distribution $p(X_a \mid X_b = x_b)$ is a Gaussian distribution $\mathcal{N}\left(\mu_{a|b}, \Sigma_{a|b}\right)$ where new mean and covariance defined by:

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

Finally plugging in the values we get

$$\mu_{a|b} = \begin{bmatrix} -0.28 \\ 2.2 \end{bmatrix}, \quad \Sigma_{aa}^{-1} = \begin{bmatrix} 0.2 & -0.9 \\ -0.9 & 4.9 \end{bmatrix}$$

Therefore, $p(x_1, x_2 \mid x_3 = 1) = \mathcal{N}\left( \left[ \begin{array}{c} x_1 \\ x_2 \end{array} \right] \middle| \begin{bmatrix} -0.28 \\ 2.2 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.9 \\ -0.9 & 4.9 \end{bmatrix} \right)$

ii. (1 point) $p(x_1 | x_3 = -1)$ and $p(x_3)$.

**Solution:** Similarly to the previous part of the question we can write that

$$p(x_1 \mid x_3 = -1) = \mathcal{N}\left(\mu_{x_1|x_3=-1}, \Sigma_{x_1|x_3=-1}\right)$$

where new mean and covariance defined by:

$$\mu_{x_1|x_3=-1} = \mu_{x_1} + \Sigma_{x_1x_3}\Sigma_{x_1x_3}^{-1}(-1-4) = -0.5$$
$$\Sigma_{x_1|x_3} = \Sigma_{x_1x_1} - \Sigma_{x_1x_3}\Sigma_{x_3x_3}^{-1}\Sigma_{x_3x_1} = 0.3 - .1 = 0.2$$

Now we can write $p(x_1 \mid x_3 = -1) = \mathcal{N}(x_1; \mu = -0.5, \Sigma = 0.2)$
The marginal distribution $p(x_3)$ follows as

$$p(x_3) = \mathcal{N}(\mu_3 = 4, \Sigma_{33} = 10)$$

(c) (4 points) In a Linear Gaussian model, we've the following Gaussian marginal distribution for $x \in \mathbb{R}^m$ and a Gaussian conditional distribution for $y \in \mathbb{R}^d$:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

We can prove that the marginal distribution of $\mathbf{y}$ is Gaussian. Given this fact, derive the mean and variance of $p(\mathbf{y})$ to show that:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top).$$

(Optional ungraded: Also show this result on the Bayes theorem equivalent for the above Linear Gaussian model: $p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\big(\mathbf{x} \,\big|\, \boldsymbol{\Sigma}\{\mathbf{A}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}\big)$, where $\boldsymbol{\Sigma} = \big(\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A}\big)^{-1}$)

**Solution:** To find the mean of the marginal $\mathbf{y}$ using total expectation we get

$$\mathrm{E}[\mathbf{y}] = \mathrm{E}_{\mathbf{x}}\left[\mathrm{E}_{\mathbf{y}|\mathbf{x}}[\mathbf{y}]\right] = \mathrm{E}_{\mathbf{x}}\left[\mathbf{A}\mathbf{x} + \mathbf{b}\right] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

Now by total covariance we write

$$\begin{aligned}\mathrm{Cov}_{\mathbf{y}}(\mathbf{y}) &= \mathrm{E}_{\mathbf{x}}\left[\mathrm{Cov}(\mathbf{y} \mid \mathbf{x})\right] + \mathrm{Cov}_{\mathbf{x}}\left(\mathrm{E}\left[\mathbf{y} \mid \mathbf{x}\right]\right) \\ &= \mathrm{E}_{\mathbf{x}}\left[\mathbf{L}^{-1}\right] + \mathrm{Cov}_{\mathbf{x}}\left(\mathbf{A}\mathbf{x} + \mathbf{b}\right) \\ &= \mathbf{L}^{-1} + \mathbf{A}\mathrm{Cov}_{\mathbf{x}}\mathbf{x}\mathbf{A}^\top = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\end{aligned}$$

2. (12 points) [LINEAR ALGEBRA + OPTIMIZATION]

(a) (4 points) Let $A \in \mathbb{R}^{m \times n}$ and let $b \notin \mathrm{ColSpace}(A)$. Then we know that the system $Ax = b$ does not admit any solution for $x$. Now:

i. (2 points) Prove that $A^\top Ax = A^\top b$ always admits a solution.
(**Hint:** Show that $\mathrm{NullSpace}(A) = \mathrm{NullSpace}(A^\top A)$).

**Solution:** If $z \in \text{Null}(A)$ then $Az = 0$. Hence

$$A^T Az = A^T(Az) = A^T 0 = 0,$$

so $z \in \text{Null}(A^T A)$. Conversely, if $z \in \text{Null}(A^T A)$ then

$$0 = z^T(A^T A)z = (Az)^T(Az) = \|Az\|^2 \Rightarrow Az = 0$$

Therefore $z \in \text{Null}(A)$. This proves $\text{Null}(A) = \text{Null}(A^T A)$.
From this equality of nullspaces we get $\text{rank}(A) = \text{rank}(A^T A)$. Hence the column spaces satisfy

$$\text{Col}(A^T A) = \text{Col}(A^T),$$

because $A^T A$ and $A^T$ have the same rank and $\text{Col}(A^T A) \subseteq \text{Col}(A^T)$ always holds. Now for any $b \in \mathbb{R}^m$ we have

$$A^T b \in \text{Col}(A^T) = \text{Col}(A^T A).$$

Hence, the linear system $A^T Ax = A^T b$ always admits a solution.

ii. (2 points) Prove that the minimizer $x^*$ of $\|Ax - b\|^2$ satisfies the equation $A^T Ax^* = A^T b$.

**Solution:** To minimize $\|Ax - b\|^2$ taking gradient w.r.to $x$ and equate to zero, we get

$$\begin{aligned}
&\frac{\partial}{\partial x}\left(\|Ax - b\|^2\right) = 0 \Rightarrow \frac{\partial}{\partial x}\left((Ax - b)^T (Ax - b)\right) = 0 \\
&\Rightarrow \frac{\partial}{\partial x}\left(x^T A^T Ax - x^T A^T b - b^T Ax + b^T b\right) = 0 \Rightarrow 2A^T Ax - 2A^T b = 0 \\
&\Rightarrow A^T Ax^* = A^T b
\end{aligned}$$

Hence, the minimizer $x^*$ indeed satisfies the equation $A^T Ax^* = A^T b$

(b) (6 points) Answer the following questions about convexity.

i. (1 point) Is $f(x) = e^{-x}$ for any $x \in \mathbb{R}$ convex? Why? What is the minima of this function?

**Solution:** $f(x) = e^{-x}$ is convex for any $x \in \mathbb{R}$. To show convexity

$$\begin{aligned}
e^{-(\lambda x + (1-\lambda)y)} = (e^{-x})^{\lambda} (e^{-y})^{1-\lambda} &\leqslant \lambda e^{-x} + (1-\lambda)e^{-y} \ [\text{ AM-GM inequality}] \\
\Rightarrow e^{-(\lambda x + (1-\lambda)y)} &\leqslant \lambda e^{-x} + (1-\lambda)e^{-y}
\end{aligned}$$

where $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$.
As $e^{-x}$ satisfies the Jensen's inequality. It is indeed convex. This function does not have a minima.

ii. (1 point) Show that the function $\Psi(t) = \log(1 + exp(-t))$ is convex.

**Solution:** Let $\lambda \in [0, 1]$ and $x, y \in \mathbb{R}$. For vectors $u = (0, -x)$ and $v = (0, -y)$,

$$\sum_{i=1}^{2} e^{\lambda u_i + (1-\lambda)v_i} = 1 + e^{-(\lambda x + (1-\lambda)y)}$$

By Holder's inequality we can write

$$\begin{aligned}
\sum_{i=1}^{2} e^{\lambda u_i + (1-\lambda)v_i} &\leqslant \left(\sum_{i=1}^{2} e^{u_i}\right)^{\lambda} \left(\sum_{i=1}^{2} e^{v_i}\right)^{1-\lambda} = (1 + e^{-x})^{\lambda}(1 + e^{-y})^{1-\lambda} \\
\Rightarrow \log(1 + e^{-(\lambda x + (1-\lambda)y)}) &\leqslant \log\left((1 + e^{-x})^{\lambda}(1 + e^{-y})^{1-\lambda}\right) \\
&= \lambda \log(1 + e^{-x}) + (1 - \lambda)(1 + e^{-y}) \\
\Rightarrow \log(1 + e^{-(\lambda x + (1-\lambda)y)}) &\leqslant \lambda \log(1 + e^{-x}) + (1 - \lambda)(1 + e^{-y})
\end{aligned}$$

Hence the $\log(1 + e^{-t})$ is convex.

iii. (4 points) Prove that the logistic loss function below is convex in $w$.

$$L(w) = \sum_{i=1}^{n} \log\big(1 + exp(-y_i w^T x_i)\big)$$

where, $w \in \mathbb{R}^d$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$ and $n$ is the number of data points (we will encounter this loss function later in the course in Logistic Regression).
(**Hint:** Sum of convex functions is convex, and also composition of a convex function and an affine function is convex. Use these results along with convexity of $\Psi$ function from last part to prove this result. Also, visualizing the above function using simple values may provide some intuition to solve the problem.)

**Solution:** For $w \in \mathbb{R}^d$, $x_i \in \mathbb{R}^d$, and $y_i \in \{-1, +1\}$, the term $-y_i w^T x_i$ is affine in $w$, hence convex. Since the exponential function is convex and increasing, $exp(-y_i w^T x_i)$ is convex, and so is $1+exp(-y_i w^T x_i)$. Because $\log(\cdot)$ is increasing, $\log\big(1 + exp(-y_i w^T x_i)\big)$ is convex. Finally, a sum of convex functions is convex, so

$$\sum_{i=1}^{n} \log\big(1 + exp(-y_i w^T x_i)\big)$$

is convex in $w$.

(c) (2 points) Given a function $f(x) = x^T Q x + c^T x + d$, where Q is a symmetric matrix, prove the following statements:

i. Convex if and only if $Q \succeq 0$.
ii. Strictly convex if and only if $Q \succ 0$.
iii. Concave if and only if $Q \preceq 0$.
iv. Strictly concave if and only if $Q \prec 0$.

Note that the symbols such as $\succeq$ and $\succ$ above refers to the definiteness of the matrix $Q$ (psd or pd in this case respectively).

**Solution:** For the quadratic function

$$f(x) = x^\top Q x + c^\top x + d,$$

whose Hessian is

$$\nabla^2 f(x) = Q.$$

i. [$\Rightarrow$] Suppose $f(x)$ is convex. By the second-order characterization of convexity, we require

$$\nabla^2 f(x) \succeq 0 \Rightarrow Q \succeq 0.$$

[$\Leftarrow$] Conversely, if $Q \succeq 0$, then

$$\nabla^2 f(x) = Q \succeq 0,$$

which implies that $f(x)$ is convex. Hence, $f(x)$ is convex $\Leftrightarrow Q \succeq 0$.

ii. [$\Rightarrow$] Suppose $f(x)$ is strictly convex. By the second-order characterization, we must have

$$\nabla^2 f(x) \succ 0 \Rightarrow Q \succ 0.$$

[$\Leftarrow$] Conversely, if $Q \succ 0$, then

$$\nabla^2 f(x) = Q \succ 0,$$

which ensures that $f(x)$ is strictly convex. Therefore, $f(x)$ is strictly convex $\Leftrightarrow Q \succ 0$.

iii. [$\Rightarrow$] Suppose $f(x)$ is concave. By the second-order characterization of concavity, we require

$$\nabla^2 f(x) \preceq 0 \Rightarrow Q \preceq 0.$$

[$\Leftarrow$] Conversely, if $Q \preceq 0$, then

$$\nabla^2 f(x) = Q \preceq 0,$$

which implies that $f(x)$ is concave. Thus, $f(x)$ is concave $\Leftrightarrow Q \preceq 0$.

iv. [$\Rightarrow$] Suppose $f(x)$ is strictly concave. Then,

$$\nabla^2 f(x) \prec 0 \Rightarrow Q \prec 0.$$

[$\Leftarrow$] Conversely, if $Q \prec 0$, then

$$\nabla^2 f(x) = Q \prec 0,$$

which ensures that $f(x)$ is strictly concave. Therefore, $f(x)$ is strictly concave $\Leftrightarrow Q \prec 0$.

3. (10 points) [FISHERIAN MLE VS. BAYESIAN MAP]

(a) (4 points) A random variable X follows the lognormal distribution defined as follows:

$$p(x) = \frac{1}{x\sqrt{\theta_2}\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \theta_1)^2}{2\theta_2}\right), \quad x > 0$$

Derive the expressions for the maximum likelihood estimates of the parameters $\theta_1$ and $\theta_2$, given a training dataset $D = \{x_1, x_2, \ldots, x_N\}$ sampled iid from the above distribution.

**Solution:** The maximum likelihood estimate

$$\theta = \arg\max_{\theta} \prod_{i=1}^{N} p(x_i) = \arg\max_{\theta} \left[\frac{1}{(x_1 x_2 \cdots x_N)(2\pi\theta_2)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{N}(\ln x_i - \theta_1)^2}{2\theta_2}\right)\right]$$

As we know taking log of the likelihood doesn't change the optimization problem so we have,

$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \ln p(x_i) = -\frac{N}{2}\ln(2\pi\theta_2) - \ln(x_1 x_2 \cdots x_N) - \sum_{i=1}^{N} \frac{(\ln x_i - \theta_1)^2}{2\theta_2}$$

Now, we write the first order condition and get

$$\frac{\partial}{\partial\theta_1}\left(\sum_{i=1}^{N} \ln p(x_i)\right) = 0 \quad ; \quad \frac{\partial}{\partial\theta_2}\left(\sum_{i=1}^{N} \ln p(x_i)\right) = 0$$

$$\Rightarrow \theta_1 = \frac{1}{N}\sum_{i=1}^{N} \ln x_i \quad ; \quad \theta_2 = \frac{1}{N}\sum_{i=1}^{N} (\ln x_i - \theta_1)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\ln x_i - \frac{1}{N}\sum_{j=1}^{N} \ln x_j\right)^2$$

Finally for maximum likelihood estimate the expressions for the parameters are

$$\theta_1 = \frac{1}{N}\sum_{i=1}^{N} \ln x_i \; ; \;\; \theta_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\ln x_i - \frac{1}{N}\sum_{j=1}^{N} \ln x_j\right)^2$$

(b) (6 points) Logan found a mystery coin and wants to know the probability of this coin landing on heads when flipped. He models the coin toss as sampling from a Bernoulli($w$) where $w$ is the probability of heads. He flips the coin three times and the flips turned out to be heads, tails,

and heads. An oracle tells him that $w \in \{0, 0.25, 0.5, 0.75, 1\}$, and *no other values of w should be considered*.

Find the MLE and Bayesian MAP (Maximum A Posteriori) estimates of $w$. Use the following prior distribution for the MAP estimate:

$$p(w) = \begin{cases} 0.9 & \text{if } w = 0, \\ 0.04 & \text{if } w = 0.25, \\ 0.03 & \text{if } w = 0.5, \\ 0.02 & \text{if } w = 0.75, \\ 0.01 & \text{if } w = 1. \end{cases}$$

**Solution:** As the samples are from Bernoulli($w$) and we observe two heads and one trail, we have the likelihood as

$$l(w) = p(D \mid w) = w^2(1-w)$$

For $w \in \{0, 0.25, 0.5, 0.75, 1\}$ calculating the likelihood

$$l(0) = 0;\ l(0.25) = \frac{3}{64};\ l(0.5) = \frac{8}{64};\ l(0.75) = \frac{9}{64};\ l(1) = 0$$

Hence the MLE estimate for $w$ is $w_{\text{MLE}} = 0.75$
For MAP estimate we calculate the posteriori as

$$p(w \mid D) \propto p(w)p(D \mid w)$$

For $w \in \{0, 0.25, 0.5, 0.75, 1\}$ calculating the posteriori

$$p(0 \mid D) \propto 0;\ p(0.25 \mid D) \propto \frac{0.12}{64};\ p(0.5 \mid D) \propto \frac{0.24}{64};\ p(0.75 \mid D) \propto \frac{0.18}{64};\ p(1 \mid D) \propto 0$$

So, the MAP estimate for $w$ is $w_{\text{MAP}} = 0.5$

4. (10 points) [PUTTING IT TOGETHER: DENSITY ESTIMATION + DECISION THEORY]

   (a) (8 points) [Optimal Classifier by Pen/Paper] Consider the following dataset:

| **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 |

   . We would like to learn a generative model $p(x, y) = p(y)p(x|y)$ from this dataset, under the assumption that the class conditionals are Gaussian distributions with a known variance of 1 and unknown means (to be estimated from the data). Also, to make optimal class assignment

for a new data point $x$, let us minimize the expected loss based on this loss matrix:

$$L = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

, where $L_{ij}$ indicates the loss for an input $x$ with $i$ being the true class and $j$ the predicted class. Given the above setup:

i. (3 points) Derive the optimal classifier $h_{MLE}(x)$ and provide the decision boundaries/regions based on MLE estimation from data.

**Solution:** From the given data, the maximum likelihood estimates (MLE) of the parameters are:

- **Class priors:** $p(y = 1) = \frac{5}{14}, \quad p(y = 2) = \frac{4}{14}, \quad p(y = 3) = \frac{5}{14}$.
- **Class-conditional means:**

$$\mu_{x|y=1} = \frac{-2.8 - 0.7 - 2.4 - 1.3 - 3.3}{5} = -2.1,$$
$$\mu_{x|y=2} = \frac{0.4 - 0.3 + 0.8 + 1.1}{4} = 0.5,$$
$$\mu_{x|y=3} = \frac{1.5 + 0.9 + 1.8 + 2.5 + 2.6}{5} = 1.86.$$

our objective is to get the optimal classifier,

$$h_{MLE}(x) = \arg\min_{j \in \{1,2,3\}} p(y \mid x) \cdot L_{:,j}$$

if $j = 1$ we have

$$p(y \mid x) \cdot L_{:,1} = p(y = 2 \mid x) + 2p(y = 3 \mid x) \propto \left[\frac{4}{14}\mathcal{N}(x \mid 0.5, 1) + \frac{10}{14}\mathcal{N}(x \mid 1.86, 1)\right] := N_1(x)$$

similarly for $j = 2$ and $j = 3$ we have

$$p(y \mid x) \cdot L_{:,2} = p(y = 1 \mid x) + p(y = 3 \mid x) \propto \left[\frac{5}{14}\mathcal{N}(x \mid -2.1, 1) + \frac{5}{14}\mathcal{N}(x \mid 1.86, 1)\right] := N_2(x)$$

$$p(y \mid x) \cdot L_{:,3} = 2p(y = 1 \mid x) + p(y = 2 \mid x) \propto \left[\frac{10}{14}\mathcal{N}(x \mid -2.1, 1) + \frac{4}{14}\mathcal{N}(x \mid 0.5, 1)\right] := N_3(x)$$

To find the decision boundaries we know at the boundaries the risks are same(i.e., $N_i = N_j, i \neq j$). So, we can write

$$N_1(x) = N_2(x) \Rightarrow x \approx -0.748$$
$$N_1(x) = N_3(x) \Rightarrow x \approx -0.120$$
$$N_2(x) = N_3(x) \Rightarrow x \approx 1.024$$

Figure 1 shows the decision boundaries (minimizing risk under the MLE model).

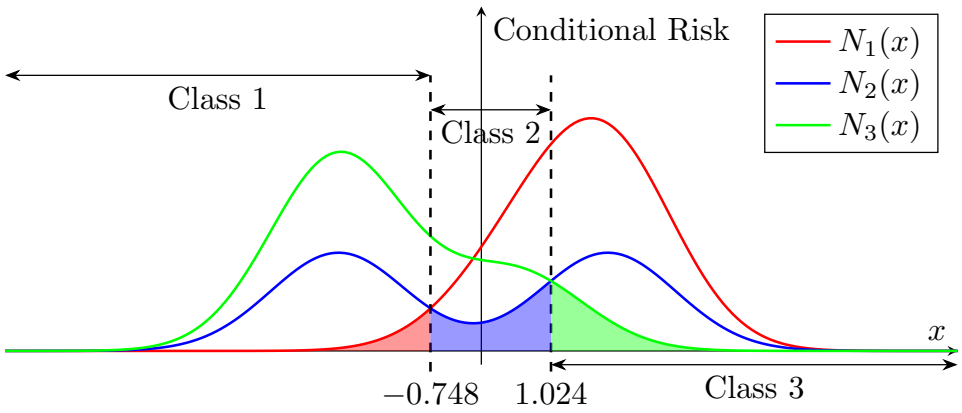


Figure 1: Conditional Risk in MLE

Thus, the optimal classifier

$$h_{MLE}(x) = \begin{cases} 1; & x \leqslant -0.748 \\ 2; & -0.748 < x \leqslant 1.024 \\ 3; & 1.024 < x \end{cases}$$

ii. (4 points) Derive the optimal classifier $h_{MAP}(x)$ and provide the decision boundaries/regions, based on the Bayesian MAP estimation from data (assuming that the parameter for the class priors follows a Dirichlet distribution with parameters/pseudocounts given by (2, 2 and 2), and the mean parameter for the class conditionals follows a standard normal distribution).

**Solution:** Given that

$$\begin{aligned} y &\sim \text{Dir}\,(\pi \mid \alpha = (2,2,2)) \\ x \mid y = 1 &\sim \mathcal{N}(\mu_1, \sigma^2 = 1); \quad \mu_1 \sim \mathcal{N}(0, \sigma_0^2 = 1) \\ x \mid y = 2 &\sim \mathcal{N}(\mu_2, \sigma^2 = 1); \quad \mu_2 \sim \mathcal{N}(0, \sigma_0^2 = 1) \\ x \mid y = 3 &\sim \mathcal{N}(\mu_3, \sigma^2 = 1); \quad \mu_3 \sim \mathcal{N}(0, \sigma_0^2 = 1) \end{aligned}$$

and by Bayesian posterior estimation we can write

$$\begin{aligned}\pi \mid D &\sim \text{Dir}(2+5, 2+4, 2+5) = \text{Dir}(7,6,7)\\ \mu_1 \mid D &\sim \mathcal{N}(\mu_{N_1}, \sigma^2_{N_1})\\ \mu_2 \mid D &\sim \mathcal{N}(\mu_{N_2}, \sigma^2_{N_2})\\ \mu_3 \mid D &\sim \mathcal{N}(\mu_{N_3}, \sigma^2_{N_3})\end{aligned}$$

where

$$\pi_1 = \frac{7-1}{7+6+7-3} = \frac{6}{17}; \quad \pi_2 = \frac{6-1}{7+6+7-3} = \frac{5}{17}; \quad \pi_3 = \frac{7-1}{7+6+7-3} = \frac{6}{17}$$

$$\mu_{N_1} = \frac{N_1\sigma_0^2}{N_1\sigma_0^2+\sigma^2}\mu_{1_{ML}} = \frac{5}{6}\times(-2.1) = -1.75; \qquad \sigma^2_{N_1} = 1/\left(\frac{N_1}{\sigma^2}+\frac{1}{\sigma_0^2}\right) = 1/6$$

$$\mu_{N_2} = \frac{N_2\sigma_0^2}{N_2\sigma_0^2+\sigma^2}\mu_{2_{ML}} = \frac{4}{5}\times(0.5) = 0.4; \qquad \sigma^2_{N_2} = 1/\left(\frac{N_2}{\sigma^2}+\frac{1}{\sigma_0^2}\right) = 1/5$$

$$\mu_{N_3} = \frac{N_3\sigma_0^2}{N_3\sigma_0^2+\sigma^2}\mu_{3_{ML}} = \frac{5}{6}\times(1.86) = 1.55; \qquad \sigma^2_{N_3} = 1/\left(\frac{N_3}{\sigma^2}+\frac{1}{\sigma_0^2}\right) = 1/6$$

Hence, the class prior are

$$p(y=1) = \frac{6}{17}; \quad p(y=2) = \frac{5}{17}; \quad p(y=3) = \frac{6}{17}$$

and the class conditional are

$$p(x \mid y=1) = \mathcal{N}(-1.75, 1)\,;\ p(x \mid y=2) = \mathcal{N}(0.4, 1)\,;\ p(x \mid y=3) = \mathcal{N}(1.55, 1)$$

our objective to get the optimal classifier

$$h_{MAP}(x) = \underset{j\in\{1,2,3\}}{\arg\min}\, p(y \mid x)\cdot L_{:,j}$$

if $j=1$ we have

$$p(y \mid x)\cdot L_{:,1} = p(y=2 \mid x) + 2p(y=3 \mid x) \propto \left[\frac{5}{17}\mathcal{N}(0.4,1) + \frac{12}{17}\mathcal{N}(1.55,1)\right] := N_1(x)$$

similarly for $j=2$ and $j=3$ we have

$$p(y \mid x)\cdot L_{:,2} = p(y=1 \mid x) + p(y=3 \mid x) \propto \left[\frac{6}{17}\mathcal{N}(-1.75,1) + \frac{6}{17}\mathcal{N}(1.55,1)\right] := N_2(x)$$

$$p(y \mid x)\cdot L_{:,3} = 2p(y=1 \mid x) + p(y=2 \mid x) \propto \left[\frac{12}{17}\mathcal{N}(-1.75,1) + \frac{5}{17}\mathcal{N}(0.4,1)\right] := N_3(x)$$

To find the decision boundaries we know at the boundaries the risks are same(i.e., $N_i = N_j, i \neq j$). So, we can write

$$N_1(x) = N_2(x) \Rightarrow x \approx -0.668$$
$$N_1(x) = N_3(x) \Rightarrow x \approx -0.1$$
$$N_2(x) = N_3(x) \Rightarrow x \approx 0.854$$

Figure 2 shows the decision boundaries (minimizing risk under the MAP model).

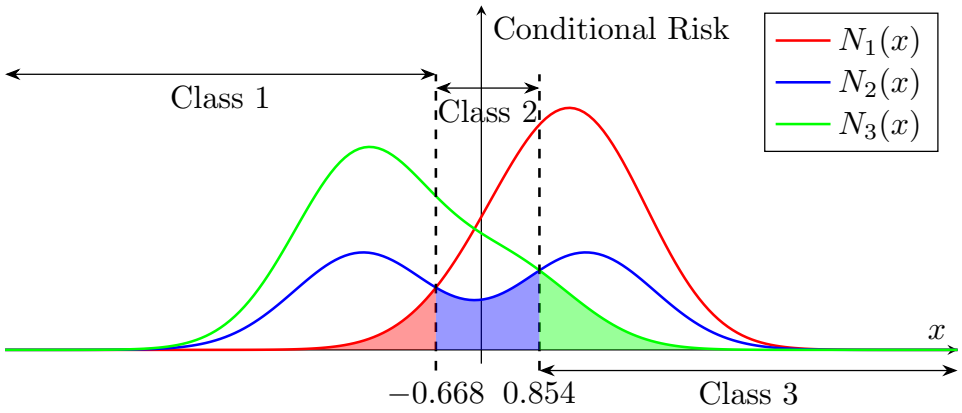


Figure 2: Conditional Risk in MAP

Thus, the optimal classifier

$$h_{MAP}(x) = \begin{cases} 1; & x \leqslant -0.668 \\ 2; & -0.668 < x \leqslant 0.854 \\ 3; & 0.854 < x \end{cases}$$

iii. (1 point) Is there any $x \in \mathbb{R}$ that is differently classified by the above two classifiers?

**Solution:** Yes there exist $x \in \mathbb{R}$ that is differently classified by the above two classifiers. For $x \in [-0.748, -0.668]$ the MLE model classifies as class 2 whereas MAP model classifies as class 1. Similarly for $x \in [0.854, 1.024]$ the MLE model classifies as class 2, where as the MAP model classifies as class 3.

(b) (2 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 $\psi$. 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:** For a test point $x$, the conditional risk of assigning it to class $C_j$ is

$$R(j \mid x) = p(y \mid x) \cdot L_{:,j}$$

And if we choose the reject option, the risk is

$$R(\text{ reject } \mid x) = \psi$$

So, the optimal classifier is

$$h(x) = \begin{cases} \arg\min_j R(j \mid x); & \min_j R(j \mid x) \leqslant \psi \\ \text{reject;} & \text{otherwise} \end{cases}$$

for $0-1$ loss we have $L_{kj} = \mathbb{1}_{k \neq j}$ then the risk becomes

$$R(j \mid x) = \sum_{k \neq j} p(C_k \mid x) = 1 - p(C_j \mid x)$$

Thus the optimal classifier is

$$\begin{aligned} h(x) &= \begin{cases} \arg\min_j R(j \mid x); & \min_j R(j \mid x) \leqslant \psi \\ \text{reject;} & \text{otherwise} \end{cases} \\ &= \begin{cases} \arg\max_j p(j \mid x); & \max_j p(j \mid x) \geqslant 1 - \psi \\ \text{reject;} & \text{otherwise} \end{cases} \end{aligned}$$

5. (15 points) [LET'S ROLL UP YOUR CODING SLEEVES...] (Note: You should follow the "General Instructions" above on how to submit your python notebook with output/results, as well as the code source files, to get full credit for this programming question.)
Given a feature vector $x \in \mathbb{R}^d$ and class labels $C \in \{c_1, \ldots, c_k\}$, the Bayes classifier assigns $x$ to the class with maximum posterior probability:

$$h(x) = \arg \max_{c \in \{c_1, \ldots, c_k\}} P(C = c \mid X = x).$$

The Naive Bayes classifier assumes conditional independence of features given the class, i.e.

$$P(X = x \mid C = c) = \prod_{j=1}^{d} P(X_j = x_j \mid C = c),$$

You are supposed to build Bayes classifiers that model each class using multivariate Gaussian density functions for the datasets assigned to you (under assumptions below and employing MLE

approach to estimate class prior/conditional densities). This assignment is focused on handling and analyzing data using interpretable classification models, rather than aiming solely for the best classification accuracy.

Build classification models for all these Case numbers (you may refer to the Chapter 2 of the book “Pattern Classification" by David G. Stork, Peter E. Hart, and Richard O. Duda):

Case 1: Bayes classifier with the same Covariance matrix for all classes.

Case 2: Bayes classifier with different Covariance matrix across classes.

Case 3: Naive Bayes classifier with the Covariance matrix $S = \sigma^2\mathbf{I}$ same for all classes.

Case 4: Naive Bayes classifier with $S$ of the above form, but being different across classes.

Refer to the provided dataset for each group, which can be found here. Each dataset includes 2D feature vectors and their corresponding class labels. There are two different datasets available:

1. Linearly separable data.
2. Non-linearly separable data.

There are 41 folders in each dataset, but you need to look at only one folder – **the folder number assigned to you** being $\texttt{RollNo}\%41 + 1$.

**Plots/answers Required**: For your assignment, you need to provide the following plots/answers (refer to the "Sample Plots" folder: link):

(a) (4 points) The plot of Gaussian pdfs for all classes estimated using the train data (train.txt). (4 Cases $\times$ 2 Datasets = 8 plots in one page)

**Solution:**

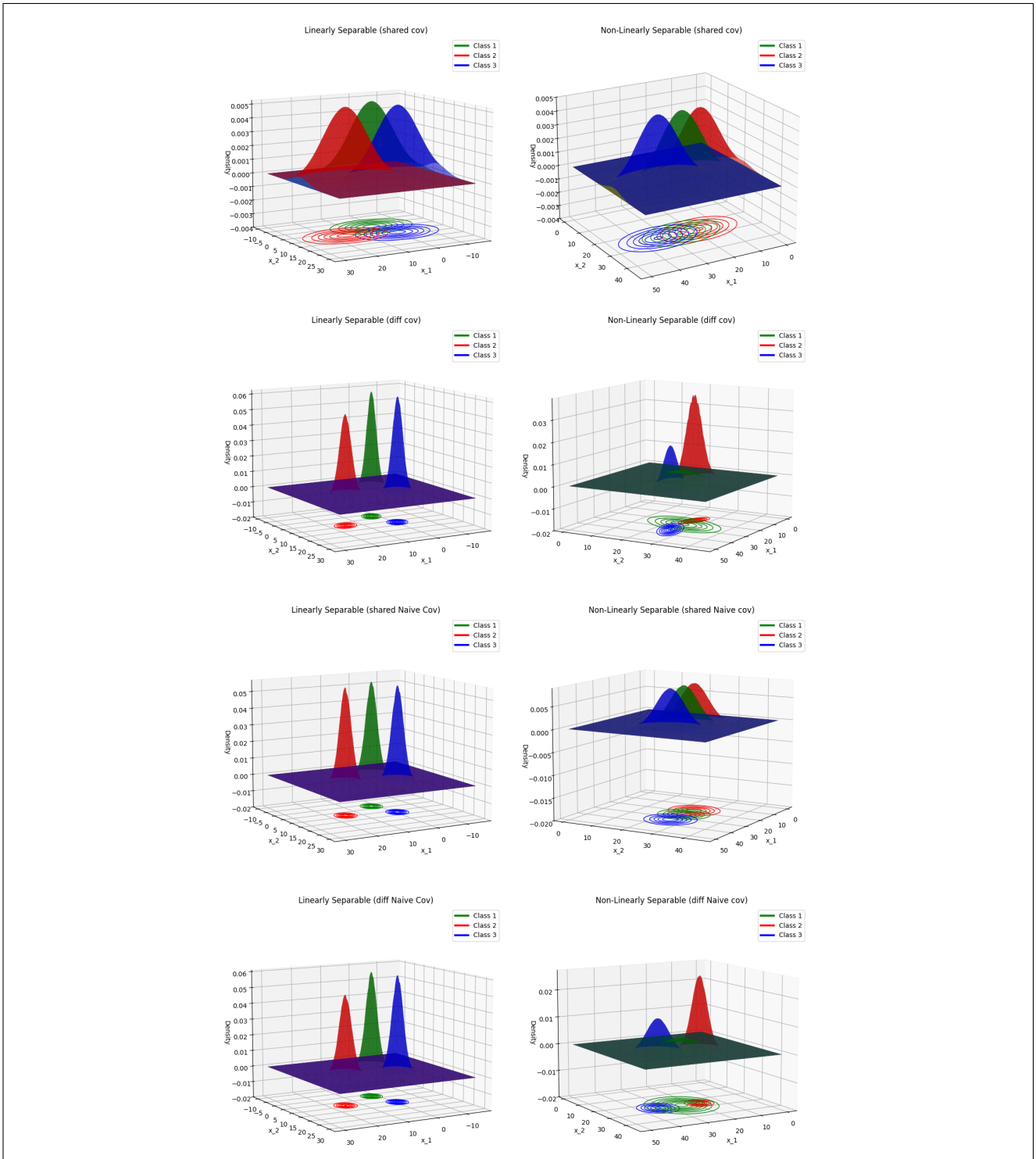


Figure 3: Gaussian pdfs

(b) (4 points) The classifiers, specifically their decision boundary/surface as a 2D plot along with training points marked in the plot (again 8 plots in one page).

**Solution:**

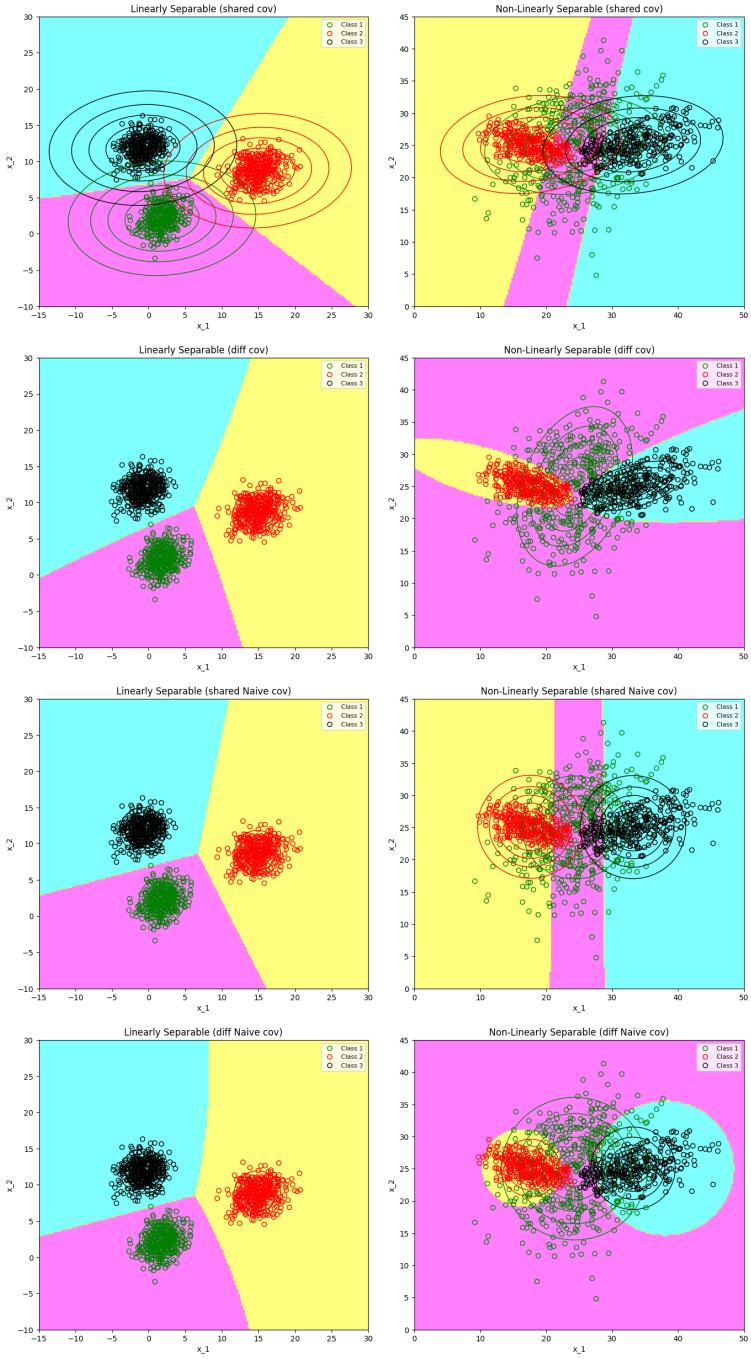


Figure 4: Decision Boundary on train data

(c) (1 point) Report the error rates for the above classifiers (four classifiers on the two datasets as a 4 × 2 table, with appropriately named rows and columns).

**Solution:** Error rate on train data case-1 & case-2:

| Bayes Classifier | Linearly Separable | Non-Linearly Separable |
|---|---|---|
| Shared Covariance | 0.0010 | 0.2438 |
| Different Covariance | 0.0010 | 0.0200 |

Error rate on train data case-3 & case-4:

| Naive Bayes Classifier | Linearly Separable | Non-Linearly Separable |
|---|---|---|
| Shared $\sigma^2$ | 0.0010 | 0.2543 |
| Different $\sigma^2$ | 0.0010 | 0.1438 |

(d) (1 point) Answer briefly on whether we can use the most general "Case 2" for all datasets? If not, answer when a simpler model like "Case 1" is preferable over "Case 2"?

**Solution:** No we can't always use the case-2. It requires estimating more parameters(i.e. different covariance matrices) and can overfit when the data is limited. Whereas case-1 is simpler mode is preferable when the class covariance are similar, data is roughly linearly separable, or/and the training set is small.

(e) (5 points) Ensure that the properly running code files that generates the above plots, etc., are submitted according to the detailed "General Instructions" in the beginning of this document.

**(Not)Allowed Libraries:** You are not allowed to use any inbuilt functions for building the model or classification using the model. However, you can use inbuilt functions/libraries for plotting and other purposes.

6. [SELF DECLARATION]
I, Ritabrata Mandal ,swear on my honour that I have prepared and written the answers for this assignment and associated code by myself and have not copied it from the internet, any LLM's output, or other students.