IITM-CS5691 : Pattern Recognition and Machine Learning — Release Date: August 30, 2023
Assignment 1 — Due Date : September 27, 2023, 23:59

Roll No: MA22C043 — Name: Shubham Singh
Collaborators (if any):
References/sources (if any): Richard O. Duda, Peter E. Hart, David G. Stork - Pattern classification (2001, Wiley)

- 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 Crowdmark 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! Instructions to join Crowdmark and submit your solution to each question within Crowdmark **TBA** 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 Crowdmark.
- 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 checks on codes).
- If you have referred a book or any other online material or LLMs (Large Language Models like ChatGPT) 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 (this also means that you cannot copy-paste the solution from LLMs!). 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 12% towards the overall course grade.

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).
   Solution:

(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 $\mathbb{R}^3$.

Graph of this approximation is $(y_1, y_2, L_v[f](y))$, this shows that it belongs to 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:

$$\begin{aligned}
P(X_1 = x_1, X_2 = x_2) &= \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(\frac{-1}{2} X^T \Sigma^{-1} X\right) \\
&= \frac{1}{2\pi\sqrt{(ad-bc)}} \exp\left(\frac{-1}{2(ad-bc)}[dx_1^2 + ax_2^2 - (c+b)x_1x_2]\right) \\
&= \frac{1}{2\pi\sqrt{(ad-bc)}} \exp\left(\frac{-d}{2(ad-bc)}[x_1^2 + \frac{a}{d}x_2^2 - \frac{(c+b)}{d}x_1x_2]\right)
\end{aligned}$$

so, $P(X_1 = x_1, X_2 = x_2) =$

$$\frac{1}{2\pi\sqrt{(ad-bc)}} \exp\left(\frac{-d}{2(ad-bc)}[x_1^2 + \frac{a}{d}x_2^2 - \frac{(c+b)}{d}x_1x_2 + \left(\frac{x_2}{2}\sqrt{\frac{b+c}{d}}\right)^2 - \left(\frac{x_2}{2}\sqrt{\frac{b+c}{d}}\right)^2]\right)$$

completing the squares,

$$\begin{aligned}
&\left[x_1^2 + \frac{a}{d}x_2^2 - \frac{(c+b)}{d}x_1x_2 + \left(\frac{x_2}{2}\sqrt{\frac{b+c}{d}}\right)^2 - \left(\frac{x_2}{2}\sqrt{\frac{b+c}{d}}\right)^2\right] \\
&\qquad = \left[\left(x_1 - \frac{x_2}{2}\sqrt{\frac{b+c}{d}}\right)^2 + x_2^2\left(\frac{a}{d} - \frac{(b+c)}{4d}\right)\right]
\end{aligned}$$

$$\begin{aligned}
P(X_1 = x_1, X_2 = x_2) &= \\
&= \sqrt{\frac{d}{2\pi(ad-bc)}} \exp\left(\frac{-d}{2(ad-bc)}\left(x_1 - \frac{x_2}{2}\sqrt{\frac{b+c}{d}}\right)^2\right) * \\
&\quad \sqrt{\frac{1}{2\pi d}} \exp\left(\frac{-d}{2(ad-bc)}x_2^2\left(\frac{a}{d} - \frac{(b+c)}{4d}\right)\right)
\end{aligned}$$

hence,

$$X_2 \sim \mathcal{N}(0, d)$$

$$X_1|X_2 = x_2 \sim \mathcal{N}\left(\frac{x_2}{2}\sqrt{\frac{b+c}{d}}, \frac{(ad-bc)}{d}\right)$$

This shows that,

$$d > 0$$

$$(b + c) > 0$$

$$\det(\Sigma) > 0$$

and,

$$\left(\frac{-d}{2(ad - bc)}\right)\left(\frac{a}{d} - \frac{(b + c)}{4d}\right) = \frac{1}{d}$$

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2) &= \frac{1}{2\pi\sqrt{(ad - bc)}} \exp\left(\frac{-1}{2(ad - bc)} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) \\ &= \frac{1}{2\pi\sqrt{(ad - bc)}} \exp\left(\frac{-1}{2(ad - bc)}[dx_1^2 + ax_2^2 - (c + b)x_1x_2]\right) \end{aligned}$$

$$\frac{1}{2\pi\sqrt{(ad - bc)}} \exp\left(\frac{-a}{2(bd - ac)}[x_2^2 + \frac{d}{a}x_1^2 - \frac{(c + b)}{a}x_2x_1 + \left(\frac{x_1}{2}\sqrt{\frac{b + c}{a}}\right)^2 - \left(\frac{x_1}{2}\sqrt{\frac{b + c}{a}}\right)^2]\right)$$

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2) = \\ &= \sqrt{\frac{a}{2\pi(ad - bc)}} \exp\left(\frac{-a}{2(bc - ad)}\left(x_2 - \frac{x_1}{2}\sqrt{\frac{b + c}{a}}\right)^2\right) * \\ &\sqrt{\frac{1}{2\pi a}} \exp\left(\frac{-a}{2(ad - bc)}x_1^2\left(\frac{d}{a} - \frac{(b + c)}{4a}\right)\right) \end{aligned}$$

Hence,

$$X_1 \sim \mathcal{N}(0, a)$$

$$X_2|X_1 = x_1 \sim \mathcal{N}\left(\frac{x_1}{2}\sqrt{\frac{b + c}{a}}, \frac{(ad - bc)}{a}\right)$$

This shows that,

$$a > 0$$

$$(b + c) > 0$$

$$\det(\Sigma) > 0$$

and,

$$\left(\frac{-d}{2(ad - bc)}\right)\left(\frac{d}{a} - \frac{(b + c)}{4a}\right) = \frac{1}{a}$$

Solution (b):

$$f(x) = x_1^2 + x_2^2 + x_1x_2$$

$$\Delta f(x) = \begin{bmatrix} 2x_1 + x_2 \\ 2x_2 + x_1 \end{bmatrix}$$

$$\Delta_v f(x) = \begin{bmatrix} 11 \\ 13 \end{bmatrix}$$

hence, linear approximation of $f(x)$ is

$$\begin{aligned} L_v[f](y) &= f(v) - \Delta_v f(x)^T (y - v) \\ &= 118 - (11y_1 - 33 + 13y_2 + 65) \\ &= 216 - 11y_1 - 13y_2 \end{aligned}$$

Solution (c):$(i), (iii)$

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 likelihood function, denoted as $L(\theta)$, yields maximum likelihood for the true parameters (mean and variance) given it is unimodal. This implies that it is a convex-down function, and the stationary point is indeed a global maximum.

Solution(b): To derive the bias of the Maximum Likelihood Estimator (MLE) for the mean (μ), we start with the MLE for μ, which is the sample mean $\hat{\mu}$. The bias $B(\hat{\mu})$ is defined as:

$$B(\hat{\mu}) = E(\hat{\mu}) - \mu$$

Now, let's calculate $E(\hat{\mu})$, the expected value of the MLE:

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Where X_i represents the individual random variables in the sample. Assuming X_i follows the true distribution with mean μ and variance σ^2, we have:

$$E(X_i) = \mu$$

Therefore,

$$E(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Finally, we can calculate the bias as:

$$B(\hat{\mu}) = \mu - \mu = 0$$

Hence, the MLE for the mean (μ) is an unbiased estimator.

To derive the bias of the Maximum Likelihood Estimator (MLE) for the variance (σ^2), we start with the MLE for σ^2, denoted as $\hat{\sigma}^2$. The bias $B(\hat{\sigma}^2)$ is defined as:

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

Now, let's calculate $E(\hat{\sigma}^2)$, the expected value of the MLE:

$$E(\hat{\sigma}^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right)$$

$$E(\hat{\sigma}^2) = \frac{1}{n} E\left(\sum_{i=1}^{n} (X_i^2 + \hat{\mu}^2 - 2X_i\hat{\mu})\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \hat{\mu}^2 - \sum_{i=1}^{n} 2X_i\hat{\mu}\right)$$

Where X_i represents the individual random variables in the sample. Assuming X_i follows the true distribution with mean μ and variance σ^2,

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^2 + n\hat{\mu}^2 - 2\hat{\mu}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^2 + n\hat{\mu}^2 - 2n\hat{\mu}^2\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^2 - n\hat{\mu}^2\right)$$

now, Var($\hat{\mu}$) = $E(\hat{\mu}^2)$ - $E(\hat{\mu})^2$

hence, $E(\hat{\mu}^2) = \frac{\sigma^2}{n} + \mu^2$

similarily, $E(X_i^2) = \sigma^2 + \mu^2$

$$E(\hat{\sigma}^2) = \frac{1}{n}\left((n\sigma^2 + n\mu^2) - n(\frac{\sigma^2}{n} + \mu^2)\right)$$

$$E(\hat{\sigma}^2) = \frac{1}{n}\left(n\sigma^2 - \sigma^2)\right)$$

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\left(\sigma^2)\right)$$

Hence, the MLE for the variance (σ) is an biased estimator.

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).

$$\mu_1 = -2.1$$
$$\mu_2 = 1$$
$$\mu_3 = 1.86$$
$$P(Y_1) = \frac{5}{14}$$
$$P(Y_2) = \frac{4}{14}$$
$$P(Y_3) = \frac{5}{14}$$
$$P(Y_1|X) = \frac{P(X|Y_1)P(Y_1)}{\sum_{i=1}^{3} P(X|Y_i)P(Y_i)}$$

$$= \frac{\frac{5}{14}\exp\left(\frac{-1}{2}(x+2.1)^2\right)}{\frac{5}{14}\exp\left(\frac{-1}{2}(x+2.1)^2\right) + \frac{4}{14}\exp\left(\frac{-1}{2}(x-1)^2\right) + \frac{5}{14}\exp\left(\frac{-1}{2}(x-1.86)^2\right)}$$

h(x) = 1, if

$$\begin{pmatrix} P(Y_1|X) \\ P(Y_2|X) \\ P(Y_3|X) \end{pmatrix} . \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \leqslant \begin{pmatrix} P(Y_1|X) \\ P(Y_2|X) \\ P(Y_3|X) \end{pmatrix} . \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

and

$$\begin{pmatrix} P(Y_1|X) \\ P(Y_2|X) \\ P(Y_3|X) \end{pmatrix} . \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \leqslant \begin{pmatrix} P(Y_1|X) \\ P(Y_2|X) \\ P(Y_3|X) \end{pmatrix} . \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

$$\implies 4\exp\left(\frac{-1}{2}(x-1)^2\right) + 5\exp\left(\frac{-1}{2}(x-1.86)^2\right) - 5\exp\left(\frac{-1}{2}(x-1.86)^2\right) \leqslant 0$$

and,

$$-1\exp\left(\frac{-1}{2}(x-1)^2\right) + 10\exp\left(\frac{-1}{2}(x-1.86)^2\right) - 5\exp\left(\frac{-1}{2}(x-1.86)^2\right) \leqslant 0$$

$\implies x < -0.5445$ and $x < -0.16804$
hence, x should be $\leqslant$-0.5445
similarliy,

$$h(x) = \begin{cases} 1, & \text{if } x < -0.5445 \\ 2, & \text{if } -0.5445 \leqslant x \leqslant 1.689 \\ 3, & \text{if } x > 1.689 \end{cases}$$

Solution (b) suppose, $P(Y_1|X) = p_1$, $P(Y_2|X) = p_2$

$$LossMatrix = \begin{pmatrix} a & b & \psi \\ c & d & \psi \end{pmatrix}$$

loss incurred for choosing h(x) =1, is

$$\begin{pmatrix} a \\ c \end{pmatrix}^T \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$$

and is has to be minimum. i.e,

$$(p_1 a + p_2 c) \leqslant (p_1 b + p_2 d)$$

and

$$(p_1 a + p_2 c) \leqslant \psi(p_1 + p_2)$$

$\implies$

$$p_1(a - b) + p_2(c - d) \leqslant 0$$

and

$$p_1(a - \psi) + p_2(c - \psi) \leqslant 0$$

similiarly, h(x) = 2

$$p_1(a - b) + p_2(c - d) \geqslant 0$$

and

$$p_1(b - \psi) + p_2(d - \psi) \leqslant 0$$

$$h(x) = \begin{cases} 1, & \text{if } p_1(a - b) + p_2(c - d) \leqslant 0, p_1(a - \psi) + p_2(c - \psi) \leqslant 0 \\ 2, & \text{if } p_1(a - b) + p_2(c - d) \geqslant 0, p_1(b - \psi) + p_2(d - \psi) \leqslant 0 \\ \psi, & \text{otherwise} \end{cases}$$

for 0-1 Loss Matrix,

$$h(x) = \begin{cases} 1, & \text{if } p_1 \geqslant p_2, p_2(1 - \psi) \leqslant p_1\psi \\ 2, & \text{if } p_1 \leqslant p_2 0, p_1(1 - \psi) \leqslant p_2\psi \\ \psi, & \text{if } otherwise \end{cases}$$

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

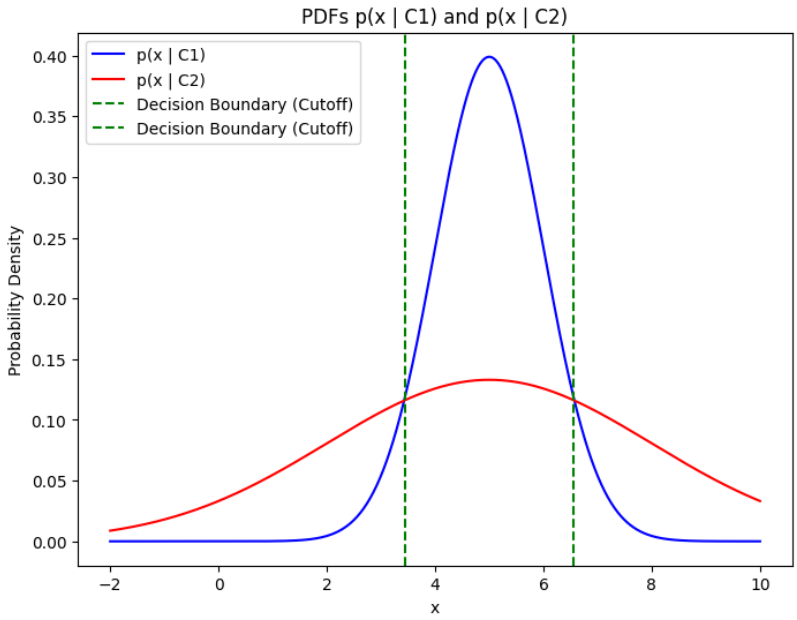
   (a) (2 points) For a two-class optimal Bayes classifer h, the decision region is given by: $R_i = \{x \in \mathbb{R} : h(x) = C_i\}$. Is R_1 always a single interval (based on a single cutoff separating the C_1 and C_2 class) or can R_1 be composed of more than one discontiguous interval? If yes for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



Solution (b):

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

$$C_{train} = \begin{bmatrix} 100 & 10 \\ 20 & 120 \end{bmatrix}$$

we know,

$$E[L] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[L_{i,j}]$$

hence,

$$= \frac{100p + 10q + 20r + 120s}{250}$$

for,

$$C_{test} = \begin{bmatrix} 90 & 45 \\ 30 & 85 \end{bmatrix}$$

$$E[L] = \frac{90p + 45q + 30r + 85s}{250}$$

Solution (c): (i): Assume $1 = positive(ill)$ and $2 = negative(healthy)$ and using MLE,

$$P(y = 1) = \frac{1}{2}$$

$$P(y = 2) = \frac{1}{2}$$
$$P(X|y = 1) \sim Bern(\theta_1) \times Bern(\theta_2) \times Bern(\theta_3)$$
$$P(X|y = 2) \sim Bern(\tau_1) \times Bern(\tau_2) \times Bern(\tau_3)$$

here,

$$\theta_1 = \tfrac{2}{3}\ \theta_1 = \tfrac{2}{3}\ \theta_1 = \tfrac{2}{3}$$
$$\tau_1 = \tfrac{1}{3}\ \tau_1 = \tfrac{1}{3}\ \tau_1 = \tfrac{1}{3}$$

the data point ($d_7 : N = -, C = +, R = -$) using a Naive Bayes classifier,

$$P(y = 1|(0,1,0)) = \frac{P((0,1,0)|y = 1) * P(y = 1)}{P((0,1,0)|y = 1) * P(y = 1) + P((0,1,0)|y = 2) * P(y = 2)}$$

$$P(y = 1|(0,1,0)) = \frac{\left(\frac{1}{3} * \frac{2}{3} * \frac{1}{3} * \frac{1}{2}\right)}{\left(\frac{1}{3} * \frac{2}{3} * \frac{1}{3} * \frac{1}{2}\right) + \left(\frac{2}{3} * \frac{1}{3} * \frac{2}{3} * \frac{1}{2}\right)}$$

$$P(y = 1|(0,1,0)) = \frac{\frac{1}{9}}{\frac{1}{9} + \frac{2}{9}}$$

$$P(y = 1|(0,1,0)) = \frac{1}{3}$$

so,

$$P(y = 2|(0,1,0)) = \frac{2}{3}$$
$$h(d_7) = 2$$

(ii): The Naive Bayes classifier assumptions are:
1. Independence Assumption: Features are conditionally independent given the class label.
2. Feature Distribution Assumption: Features follow specific probability distributions.
3. Class Prior Probability: The prior probabilities of classes are known and remain constant.

(iii): let $x = (x_1, x_2, x_3)$ and $\eta(x) = P(y = 1|x)$ so,

$$h(x) = \begin{cases} 1, & \text{if } \eta(x) \geqslant \frac{1}{2} \\ 2, & \text{otherwise} \end{cases}$$

(iv): where X_i are Bernuolli distribution.

5. (16 points) [LET'S ROLL UP YOUR CODING SLEEVES...] (Note: You should follow instructions in the preamble on how to submit notebook with output/results, as well as the code source files, to get full credit for this programming question.)
You are supposed to build Bayesian 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 Bayesian models for the given 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)

(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).

(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).

(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"?

(e) (6 points) Ensure that the properly running code files that generates the above plots, etc., are submitted according to the detailed instructions in the preamble.

(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.