IITM-CS5691 : Pattern Recognition and Machine Learning
Assignment 1

Release Date: March 9, 2021
Due Date : March 23, 23:59

Roll No: **EE18B035**
Name: **Nithin Uppalapati**
Collaborators (if any):
References (if any):

- Use LaTeX to write-up your solutions (in the solution blocks of the source LaTeX file of this assignment), and submit the resulting single pdf file 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! Within GradeScope, indicate the page number where your solution to each question starts, else we won't be able to grade it! You can join GradeScope using course entry code **5VDNKV**).
- 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 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 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.
- Points will be awarded based on how clear, concise and rigorous your solutions are, and how correct your code is. Overall points for this assignment would be **min**(your score including bonus points scored, 50).

1. (10 points) [GETTING YOUR BASICS RIGHT!]

   (a) (1 point) You have a jar of $1,000$ coins. 999 are fair coins, and the remaining coin will always land heads. You take a single coin out of the jar and flip it 10 times in a row, all of which land heads. What is the probability your next toss with the same coin will land heads? Explain your answer. How would you call this probability in Bayesian jargon?

Solution:

1. Let $\mathcal{A}$ be a random variable defined as:

$\mathcal{A}$ represents the type of coin drawn from the jar	
Fair Coin is drawn	Made-up Coin is drawn
(Total: 999)	(Total: 1)
$A = 1$	$A = 2$
$P(A = 1) = \frac{999}{1000}$	$P(A = 2) = \frac{1}{1000}$

2. Let $\mathcal{B}$ be a random variable defined as:

$\mathcal{B}$ represents the type of output of coin when tossed	
Heads is the result	Tails is the result
$B = 1$	$B = 2$
$P(B = 1|A = 1) = 0.5$	$P(B = 2|A = 1) = 0.5$
$P(B = 1|A = 2) = 1$	$P(B = 2|A = 2) = 0$

3. Let event $\mathcal{C}$ be defined as: tossing the coin for 11^{th} time, and landing with result as Heads.

4. Let event $\mathcal{D}$ be defined as: the first 10 consecutive tosses result (all of them) in Heads.

We need to find, $\mathcal{P}(\mathcal{C}|\mathcal{D})$:

Which is same as,

$$\mathcal{P}(\mathcal{C}|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{C}, \mathcal{D})}{\mathcal{P}(\mathcal{D})} \tag{1}$$

First, we solve $\mathcal{P}(\mathcal{D})$: which can be written as (total law of probability):

$$\mathcal{P}(\mathcal{D}) = \sum_{i=1}^{2} \mathcal{P}(\mathcal{A} = i)\mathcal{P}(\mathcal{B}_1 = 1, \mathcal{B}_2 = 1, \ldots, \mathcal{B}_{10} = 1|\mathcal{A} = i) \tag{2}$$

Where $\mathcal{B}_i$ indicates the outcome of i_{th} toss and $\mathcal{B}$=1 implies that the outcome is a Head.
Now, as the outcomes of the coins are independent (conditionally) w.r.t each other...
Hence we can write $\mathcal{P}(\mathcal{B}_1 = 1, \mathcal{B}_2 = 1, \ldots, \mathcal{B}_{10} = 1|\mathcal{A} = i)$ as $\prod_{j=1}^{10} \mathcal{P}(\mathcal{B}_j = 1|\mathcal{A} = i)$.

Now we compute $\mathcal{P}(\mathcal{D})$:

$$\mathcal{P}(\mathcal{D}) = (0.001 * \{\prod_{j=1}^{10}(1)\}) + (0.999 * \{\prod_{j=1}^{10}(0.5)\}) \tag{3}$$

Which is $\mathcal{P}(\mathcal{D}) = 0.0019755859375$

Now we compute $\mathcal{P}(\mathcal{C}, \mathcal{D})$:

$$\mathcal{P}(\mathcal{C}, \mathcal{D}) = (0.001 * \{\prod_{j=1}^{11}(1)\}) + (0.999 * \{\prod_{j=1}^{11}(0.5)\}) \tag{4}$$

Which is $\mathcal{P}(\mathcal{C}, \mathcal{D}) = 0.00148779296875$

Hence, $\mathcal{P}(\mathcal{C}|\mathcal{D}) = \frac{0.00148779296875}{0.0019755859375} = 0.7530894$.

If this were Bayesian Setting, as were are calculating the probability from the joint probability and dividing it by evidence, we call it as Posterior Probability.

(b) (3 points) 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. How can you argue that the stationary points so obtained are indeed global maxima of the likelihood function? Next, derive the bias of the MLE of μ, σ^2.

Solution: Given that the $x_i's$ are sampled from I.I.D data.
First, we compute likelihood function.

$$\mathcal{L}(\theta; \mathcal{D}) = \mathcal{P}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \mathcal{P}(x_i; \theta), \qquad \text{(under iid assumption)}$$

Where $\mathcal{P}(x_i)$ is given by:

$$\mathcal{P}(x_i; \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{(2\sigma^2)}} \qquad \text{where } \theta = \{\mu, \sigma\}$$

Now the choose the θ in such a way that it maximises the **Likelihood** function.

$$\hat{\theta}(\mathbf{X}) = \arg\max_\theta \mathcal{L}(\theta \mid \mathbf{X})$$

Now compute the likelihood function:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{(2\sigma^2)}} \tag{5}$$

Now apply Log to both sides of the equation 5. We get,

$$\ln\left(\mathcal{L}(\theta; \mathcal{D})\right) = \sum_{i=1}^{n} \left(-\ln(\sigma\sqrt{2\pi}) - \frac{(x_i - \mu)^2}{(2\sigma^2)} \right) \tag{6}$$

Now differentiating the equation 6 w.r.t θ, i.e., w.r.t μ and σ, we get:

$$\frac{\partial \ln(\mathcal{L}(\theta;\mathcal{D}))}{\partial \mu} = 0$$
$$\sum_{i=1}^{n} \frac{2(x_i-\mu)(-\partial\mu/\partial\mu)}{2\sigma^2} = 0$$
$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Similarly, differentiating w.r.t σ

$$\frac{\partial \ln(\mathcal{L}(\theta;\mathcal{D}))}{\partial \sigma} = 0$$
$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{\sigma} = 0$$

upon rearranging we get, $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n}$

For finding the Bias of estimators $\hat{\mu}$ and $\hat{\sigma}$,

$$\begin{aligned}
E[\hat{\mu}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} x_1\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[x_i] \quad \text{; (under iid assumption)} \\
&= \frac{1}{n} n E[x] = E[x] = \mu \implies \text{Bias}_\mu = \mu - \mu = 0
\end{aligned}$$

$$
\begin{aligned}
E\left[\hat{\sigma}^2\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{N}x_i^2 - 2\sum_{i=1}^{n}x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}\left(x_i^2-\mu^2\right) - n\left(\bar{x}^2-\mu^2\right)\right] \text{(; adding and subtracting } n\mu^2\text{)} \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}\left(x_i^2-\mu^2\right)\right] - E\left[\bar{x}^2-\mu^2\right] \\
&= \frac{n}{n}E\left[x^2-\mu^2\right] - V\left[\bar{x}\right]\text{; (under iid assumption)} \\
&= V\left[x\right] - V\left[\bar{x}\right] \\
&= V\left[x\right] - V\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right] \\
&= V\left[x\right] - \frac{1}{n^2}V\left[\sum_{i=1}^{n}x_i\right] \\
&= V\left[x\right] - \frac{n}{n^2}V\left[x\right]\text{; (under iid assumption)} = \sigma_x^2(1-\frac{1}{n})
\end{aligned}
$$

$\implies$ $\text{Bias}_\sigma = \frac{-1}{n}\sigma_x^2$

To prove that the stationary points so obtained are indeed global maxima of the likelihood function, we can check the second derivative.

At the MLE point, i.e. $\hat{\mu} = \sum_{i=1}^{n}x_i/n$, we get:

$\frac{\partial^2}{\partial\mu^2}\ln\mathcal{P} = -\frac{n}{\sigma^2}$

Now, we find the second derivative w.r.t $\ln\sigma$ at the estimated value of σ_x: we get,

$\frac{\partial^2}{\partial(\ln(\sigma))^2}\ln(\mathcal{P}) = -2n$

So, both the second derivatives are negative, and also there is a single root for the first order derivatives. Hence the ML estimate which is obtained is a stationary point.

(c) (2 points) Consider a hyperplane $\mathbb{H}$ in $\mathbb{R}^d$ passing through zero. Prove that $\mathbb{H}$ is a subspace of $\mathbb{R}^d$ and is of dimension $d-1$.

Solution: Let the equation of the hyperplane be $\mathbf{a}^T \cdot \mathbf{x} = 0$. **a** is a non-zero vector, **a** and **x** are "d" dimensional vectors. **a** is the vector normal to the hyperplane, and x is any point which lies on the hyperplane.
Now consider the vector a and extend the basis, especially, extend the basis such that it becomes a orthonormal basis (Ex: by Gram-Schmidt orthonormalization method). As the vectors in the extended basis are orthogonal to each other, and also the number of vectors in this basis is equal to d.
Now any vector in the $\mathbb{R}^d$ can be represented in terms of these basis. Lets consider an arbitrary vector **y**. It can be represented as,

$$\mathbf{y} = k_0\mathbf{a} + k_1\mathbf{e_1} + \ldots + k_{d-1}\mathbf{e_{d-1}}$$

For y to be on the Hyperplane, it should satisfy: $\mathbf{a}^T \cdot \mathbf{y} = 0$. Which means,

$$\mathbf{a}^T \cdot \mathbf{y} = k_0\mathbf{a}^T \cdot \mathbf{a} + k_1\mathbf{a}^T \cdot \mathbf{e_1} + \ldots + k_{d-1}\mathbf{a}^T \cdot \mathbf{e_{d-1}}$$

As the basis is orthonormal, we have $\mathbf{a}^T \cdot \mathbf{e_i} = 0;\ \forall i \in \{1, \ldots, d-1\}$. So for y to lie on the hyper plane, k_0 must be zero. (Which means that component of **a** in vector **y** must be zero)
So now **y** can only be combination of the remaining $d-1$ vectors, hence the dimension of Hyperplane is given by the dimension range of the matrix $\mathcal{E} = [\mathbf{e_1}, \ldots, \mathbf{e_{d-1}}]$. According to the **Rank-Nullity** Theorem,
Rank + Nullity = d
and as the columns of $\mathcal{E}$ are linearly independent, the rank is $d-1$ (Nullity is 1, as the matrix is of size $d \times d-1$
Hence the range of matrix $\mathcal{E}$ is d-1. Hence the possible dimensions for **y** (Arbitrary vector lying on Hyperplane) is $d-1$.
So dimension of $\mathbb{H}$ is $d-1$.

(d) (2 points) We saw a mixture of two 1D Gaussians ($N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$) in class with parameters π_1, π_2 for the mixing proportions. Is the likelihood of this model convex or not convex? Give proof to support your view.

Solution:

$$\begin{aligned}
\hat{\theta}_{ML} &= \arg\max_{\theta} \left(\prod_{x \in \mathcal{D}_N} P(x;\theta) \right) \\
&= \arg\max_{\theta} \prod_{x \in \mathcal{D}_N} (P(\mathbf{x};\theta)) \\
&= \arg\max_{\theta} \prod_{x \in \mathcal{D}_N} \left(\sum_z P(\mathbf{x}, z;\theta) \right) \quad = \arg\max_{\theta} \left(\sum_z P(\mathbf{x}|z;\theta) P(z) \right)
\end{aligned} \tag{7}$$

$f : \mathbf{R}^n \to \mathbf{R}$ is convex $\forall x, y \in \mathbf{R}^n$ and $\forall \lambda \in [0, 1]$ if it satisfies:

$$f(\lambda x + (1 - \lambda y)) \leqslant \lambda f(x) + (1 - \lambda) f(y))$$

Equality holds when either λ =1 or $x = y$.
Let us consider the model: $\mathcal{L} = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \sigma_2)$, and $\pi_1 + \pi_2 = 1$.
Now the ` is given by $[\mu_1, \sigma_1, \mu_2, \sigma_2, \pi_1]$. π_2 is excluded because π_2 and π_1 are dependent on each other.

We can prove that the log likelihood need not be strictly convex, by contradiction.
Consider $\lambda = 0.5$ and assume $\theta_1 = [-0.5, 1, +0.5, 1, 0.5]$
So, the likelihood now looks like:

$$0.5\frac{1}{\sqrt{2\pi}} e^{\frac{-(x-0.5)^2}{(2)}} + 0.5\frac{1}{\sqrt{2\pi}} e^{\frac{-(x+0.5)^2}{(2)}}$$

If we take $x_1 = +2$ and $x_2 = -2$ we get; $0.5\frac{1}{\sqrt{2\pi}} e^{\frac{-(-2-0.5)^2}{(2)}} + 0.5\frac{1}{\sqrt{2\pi}} e^{\frac{-(2+0.5)^2}{(2)}}$ which is numerically equal to 0.0350. And as we chose the value of λ to be 0.5, the convex combination of x and y gives 0. When we substitute 0 in likelihood we get $0.5\frac{1}{\sqrt{2\pi}} e^{\frac{-(-0.5)^2}{(2)}} + 0.5\frac{1}{\sqrt{2\pi}} e^{\frac{-(0.5)^2}{(2)}}$ which is numerically equal to 0.7041.

As $f(\lambda x + (1 - \lambda y)) > \lambda f(x) + (1 - \lambda) f(y))$, the likelihood function is not convex.

(e) (2 points) Show that there always exists a solution for the system of equations, $A^T Ax = A^T b$, where $x \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$. Further, show that for some solution x^* of this system of equations, Ax^* is the projection of b onto the column space of A.

Solution: Consider matrix $\mathbf{A} = [\mathbf{a_1}, \ldots, \mathbf{a_m}]$; it need not be invertible. Where $\mathbf{a_i} \in \mathbb{R}^n$, $\forall\, i \in \{1, \ldots, m\}$. Now, let us define a subspace which is column space of $\mathbf{A}$,
$\mathbf{U} : \{\mathbf{a_1}, \ldots, \mathbf{a_m}\} \in \mathbf{U}$
Also now consider the orthonormal basis for the above subspace.
Which is $\{\mathbf{e_1}, \ldots, \mathbf{e_k}\} \in \mathbf{U}$, where $k \leqslant \min(n, m)$ is the Rank of matrix $\mathbf{A}$. Now extend the orthonormal basis so that we get, $\{\mathbf{e_1}, \ldots, \mathbf{e_k}, \mathbf{e_{k+1}}, \ldots, \mathbf{e_n}\}$. So now $\{\mathbf{e_{k+1}}, \ldots, \mathbf{e_n}\} \in \mathbf{U}^\perp$.
Which means: $\mathbf{a_i} \cdot \mathbf{e_j} = 0\ \forall (i \in \{1, \ldots, m\}; j \in \{k + 1, \ldots, n\})$
Now, consider an arbitrary vector $\mathbf{b} \in \mathbb{R}^n$. So $\mathbf{b}$ can be orthogonally decomposed as follows:
$\mathbf{b} = \mathbf{b_u} + \mathbf{b_{u^\perp}}$.
$\mathbf{b_u}$ and $\mathbf{b_{u^\perp}}$ can be found as follows:
$\mathbf{b_u} = \sum_{i=1}^{k} \langle \mathbf{b}, \mathbf{e_i} \rangle \mathbf{e_i}$ and similarly,

$\mathbf{b_{u^\perp}} = \sum_{j=k+1}^{n} \langle \mathbf{b}, \mathbf{e_j} \rangle \mathbf{e_j}$.
As, $\mathbf{b_u}$ is a linear combination of vectors $\{\mathbf{e_1}, \ldots, \mathbf{e_k}\} \in \mathbf{U}$, thus it belongs to the subspace $\mathbf{U}$. Hence $\mathbf{b_u}$ belongs to columnspace of matrix $\mathbf{A}$. So, always $\exists\ \mathbf{x}^* : \mathbf{A}.\mathbf{x}^* = \mathbf{b_u}$.

And also we can show that $\mathbf{b_u}$ is the orthogonal projection of $\mathbf{b}$ on the subspace $\mathbf{U}$.

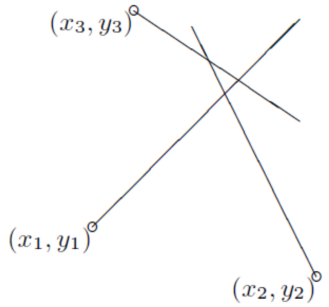
$$\begin{aligned}
\mathbf{b} &= \mathbf{b_u} + \mathbf{b_{u^\perp}} \\
\mathbf{b} - \mathbf{b_u} &= \mathbf{b_{u^\perp}} \in \mathbf{U}^\perp \\
\mathbf{U} \cdot (\mathbf{b} - \mathbf{b_u}) &= \mathbf{U} \cdot \mathbf{b_{u^\perp}} = 0 \text{ , as } \mathbf{U} \text{ and } \mathbf{U}^\perp \text{ are orthogonal}
\end{aligned}$$

Now we consider $\mathbf{A}.\mathbf{x}^* - \mathbf{b}$:

$$\begin{aligned}
\mathbf{A}.\mathbf{x}^* - \mathbf{b_u} - \mathbf{b_{u^\perp}} &= 0 - \mathbf{b_{u^\perp}} \in \mathbf{U}^\perp \\
\mathbf{u}^\mathsf{T} \cdot \mathbf{U}^\perp &= 0 \text{ where } \mathbf{u} \text{ belongs to U} \\
\mathbf{u}^\mathsf{T} \cdot \mathbf{b_{u^\perp}} &= 0 \\
\implies \mathbf{u}^\mathsf{T} \cdot (\mathbf{A}.\mathbf{x}^* - \mathbf{b}) &= 0 \\
\begin{bmatrix} \mathbf{a_1}^\mathsf{T} \\ \vdots \\ \mathbf{a_m}^\mathsf{T} \end{bmatrix} \cdot (\mathbf{A}.\mathbf{x}^* - \mathbf{b}) &= \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \\
\implies \mathbf{A}^\mathsf{T} \cdot (\mathbf{A}.\mathbf{x}^* - \mathbf{b}) &= 0
\end{aligned}$$

Hence, Proved.

2. (5 points) [OF SAILORS AND BEARINGS...] A sailor infers his location (x, y) by measuring the bearings of three buoys whose locations (x_n, y_n) are given on his chart. Let the true bearings of the buoys be θ_n (measured from north as explained here). Assuming that his measurement $\tilde{\theta}_n$ of each bearing is subject to Gaussian noise of small standard deviation σ, what is his inferred location, by maximum likelihood?



The sailor's rule of thumb says that the boat's position can be taken to be the centre of the cocked hat, the triangle produced by the intersection of the three measured bearings as in the figure shown. Can you persuade him that the maximum likelihood answer is better?

Solution:

3. (5 points) [REVEREND BAYES DECIDES]

   (a) (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 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 $L_{kj} = 1 - I_{kj}$, with I_{kj} being 1 for $k = j$ and 0 otherwise).

   Solution: Loss incurred when we make a decision as class $\mathcal{C}_j$ is :

   $$\mathbb{L}_j \equiv \sum_k [L_{kj}.p(x, \mathcal{C}_k)]$$

   Whereas the loss incurred when we not choose to predict anything is : ψ. So, in order to make optimal decision for class $\mathcal{C}_j$ it must satisfy below criteria:

   $$\text{Choose class } \mathcal{C}_j \text{ if, } \quad \mathbb{L}_j \leqslant \min[\psi,\ \mathbb{L}_i : i \neq j]$$

   $$\text{We choose to reject if, } \psi \leqslant \min[\ \mathbb{L}_i : \forall i\ ]$$

   Solving for x which satisfies the above set of inequalities, gives us the decision region for class $\mathcal{C}_j$.

   (b) (2 points) 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. All the three classes are equally likely to occur. The class densities are $P(x|C_1) \sim N(-2, 1)$, $P(x|C_2) \sim N(0, 1)$ and $P(x|C_3) \sim N(2, 1)$. Find the Bayes classifier $h(x)$.

   Solution: The overall loss incurred is:

   $$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p\left(\mathbf{x}, \mathcal{C}_k\right) d\mathbf{x}$$

We define $\mathcal{R}_k$ such that , if $\mathbf{x} \in \mathcal{R}_k$, we say that $h(\mathbf{x}) = \mathcal{C}_k$; $\forall \mathbf{x} \in \mathcal{R}_k$. Where $h(\mathbf{x})$ is the predicted value for a given data point $\mathbf{x}$. Given that $p(\mathbf{x}, \mathcal{C}_1) = p(\mathbf{x}, \mathcal{C}_2) = p(\mathbf{x}, \mathcal{C}_3) = \frac{1}{3}$
The loss incurred when we assign $h(\mathbf{x}) = \mathcal{C}_k$; when $\mathbf{x} \in \mathcal{R}_k$ is given as,

$$\int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, d\mathbf{x} \tag{8}$$

In order to have a optimized prediction, we have to minimise the loss given by the expression 8. Which is same as minimising the expression 9

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) \tag{9}$$

So, in order to determine the region $\mathcal{R}_j'^s$, we have to solve for the below set of equations:

$$\begin{aligned}
\mathcal{R}_1 : \sum_{k=1}^{3} L_{k1} p(\mathbf{x}, \mathcal{C}_k) &\leqslant \operatorname{argmin}\left(\sum_{k=1}^{3} L_{k2} p(\mathbf{x}, \mathcal{C}_k), \sum_{k=1}^{3} L_{k3} p(\mathbf{x}, \mathcal{C}_k) \right) \\
\mathcal{R}_2 : \sum_{k=1}^{3} L_{k2} p(\mathbf{x}, \mathcal{C}_k) &\leqslant \operatorname{argmin}\left(\sum_{k=1}^{3} L_{k3} p(\mathbf{x}, \mathcal{C}_k), \sum_{k=1}^{3} L_{k1} p(\mathbf{x}, \mathcal{C}_k) \right) \\
\mathcal{R}_3 : \sum_{k=1}^{3} L_{k3} p(\mathbf{x}, \mathcal{C}_k) &\leqslant \operatorname{argmin}\left(\sum_{k=1}^{3} L_{k1} p(\mathbf{x}, \mathcal{C}_k), \sum_{k=1}^{3} L_{k2} p(\mathbf{x}, \mathcal{C}_k) \right)
\end{aligned}$$

For $\mathcal{R}_1$ we get:

$$\begin{aligned}
p(\mathbf{x}, \mathcal{C}_2) + 2p(\mathbf{x}, \mathcal{C}_3) \leqslant p(\mathbf{x}, \mathcal{C}_1) + p(\mathbf{x}, \mathcal{C}_3) \; ; \; p(\mathbf{x}, \mathcal{C}_2) + 2p(\mathbf{x}, \mathcal{C}_3) &\leqslant 2p(\mathbf{x}, \mathcal{C}_1) + p(\mathbf{x}, \mathcal{C}_2) \\
e^{\frac{-(x+2)^2}{2}} - e^{\frac{-(x-2)^2}{2}} \geqslant e^{\frac{-(x)^2}{2}} \; ; \; e^{\frac{-(x+2)^2}{2}} &\geqslant e^{\frac{-(x-2)^2}{2}} \\
e^{\frac{(x)^2}{2}} . e^{\frac{-(x^2+4x+4)}{2}} - e^{\frac{-(x^2-4x+4)}{2}} \geqslant 1 \; ; \; \frac{(x+2)^2}{2} &\leqslant \frac{(x-2)^2}{2} \\
\frac{1}{2}(e^{2x} - e^{-2x}) \leqslant -\frac{e^2}{2} \; ; \; (x+2)^2 &\leqslant (x-2)^2 \\
\sinh 2x \leqslant -\frac{e^2}{2} \; ; \; x &\leqslant 0 \\
x \leqslant \frac{-1}{2} \sinh^{-1}\left(\frac{e^2}{2}\right) \; ; \; x &\leqslant 0
\end{aligned}$$

Hence for $\mathcal{R}_1$ we get $x \leqslant -1.0089$

For $\mathcal{R}_2$ we get:

$$p(\mathbf{x}, \mathcal{C}_2) + 2p(\mathbf{x}, \mathcal{C}_3) \geqslant p(\mathbf{x}, \mathcal{C}_1) + p(\mathbf{x}, \mathcal{C}_3) \; ; \; p(\mathbf{x}, \mathcal{C}_2) + 2p(\mathbf{x}, \mathcal{C}_1) \geqslant p(\mathbf{x}, \mathcal{C}_1) + p(\mathbf{x}, \mathcal{C}_3)$$

$$e^{\frac{-(x+2)^2}{2}} - e^{\frac{-(x-2)^2}{2}} \leqslant e^{\frac{-(x)^2}{2}} \; ; \; e^{\frac{-(x-2)^2}{2}} - e^{\frac{-(x+2)^2}{2}} \leqslant e^{\frac{-(x)^2}{2}}$$

$$e^{\frac{(x)^2}{2}}.e^{\frac{-(x^2+4x+4)}{2}} - e^{\frac{-(x^2-4x+4)}{2}} \leqslant 1 \; ; \; e^{\frac{(x)^2}{2}}.e^{\frac{-(x^2-4x+4)}{2}} - e^{\frac{-(x^2+4x+4)}{2}} \leqslant 1$$

$$\frac{1}{2}(e^{2x} - e^{-2x}) \geqslant -\frac{e^2}{2} \; ; \; \frac{1}{2}(e^{2x} - e^{-2x}) \leqslant \frac{e^2}{2}$$

$$\sinh 2x \geqslant -\frac{e^2}{2} \; ; \; \sinh 2x \leqslant \frac{e^2}{2}$$

$$x \geqslant \frac{-1}{2}\sinh^{-1}\left(\frac{e^2}{2}\right) \; ; \; x \leqslant \frac{1}{2}\sinh^{-1}\left(\frac{e^2}{2}\right)$$

Hence for $\mathcal{R}_2$ we get $-1.0089 \leqslant x \leqslant 1.0089$

For $\mathcal{R}_3$ we get:

$$p(\mathbf{x}, \mathcal{C}_2) + 2p(\mathbf{x}, \mathcal{C}_3) \geqslant 2p(\mathbf{x}, \mathcal{C}_1) + p(\mathbf{x}, \mathcal{C}_2) \; ; \; p(\mathbf{x}, \mathcal{C}_2) + 2p(\mathbf{x}, \mathcal{C}_1) \leqslant p(\mathbf{x}, \mathcal{C}_1) + p(\mathbf{x}, \mathcal{C}_3)$$

$$x \geqslant \frac{1}{2}\sinh^{-1}\left(\frac{e^2}{2}\right) \; ; \; x \geqslant 0$$

Hence for $\mathcal{R}_3$ we get $x \geqslant +1.0089$

$$h(x) = \begin{cases} \mathcal{C}_1 & x \leqslant -1.0089 \\ \mathcal{C}_2 & -1.0089 \leqslant x \leqslant +1.0089 \\ \mathcal{C}_3 & x \geqslant +1.0089 \end{cases}$$

(c) (1 point) Consider two classes C_1 and C_2 with equal priors and with class conditional densities of a feature x given by Gaussian distributions with respective means μ_1 and μ_2, and same variance σ^2. Find equation of the decision boundary between these two classes.

Solution: As this is just a binary class problem, for a given x, we directly can say that if $p(\mathcal{C}_1|x)$ is less than $p(\mathcal{C}_2|x)$, which means the probability of x being $\mathcal{C}_2$ is greater than that of x being $\mathcal{C}_1$. Hence the optimal decision would be assigning x to $\mathcal{C}_2$.

Hence exactly at decision boundary the conditional probabilities are equal. We write,

$$\left(p(\mathcal{C}_1|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu_1)^2}{(2\sigma^2)}}\right) = \left(p(\mathcal{C}_2|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu_2)^2}{(2\sigma^2)}}\right)$$

$$e^{\frac{-(x-\mu_1)^2}{(2\sigma^2)}} = e^{\frac{-(x-\mu_2)^2}{(2\sigma^2)}}$$

$$\frac{-(x-\mu_1)^2}{(2\sigma^2)} = \frac{-(x-\mu_2)^2}{(2\sigma^2)}$$

$$(x-\mu_1)^2 = (x-\mu_2)^2 \implies (x-\mu_1)^2 - (x-\mu_2)^2 = 0$$

$$(2x-\mu_1-\mu_2)(\mu_2-\mu_1) = 0$$

If we assume that $\mu_2 \neq \mu_1$, then we get, $x = \frac{\mu_1+\mu_2}{2}$. Which is the required decision boundary.

4. (10 points) [DON'T MIX YOUR WORDS!]
Consider two documents D_1, D_2 and a background language model given by a Categorical distribution (i.e., assume $P(w|\theta)$ is known for every word w in the vocabulary V). We use the maximum likelihood method to estimate a unigram language model based on D_1, which will be denoted by θ_1 (i.e, $p(w \mid \theta_1)$ = "nos. of times word w occurred in D_1"/ $|D_1|$, where $|D_1|$ denotes the total number of words in D_1). Assume document D_2 is generated by sampling words from a two-component Categorical mixture model where one component is $p(w|\theta_1)$ and the other is $p(w|\theta)$. Let λ denote the probability that D_1 would be selected to generate a word in D_2. That makes $1-\lambda$ the probability of selecting the background model. Let $D_2 = (w_1, w_2, ...w_k)$, where w_i is a word from the vocabulary V. Use the mixture model to fit D_2 and compute the ML estimate of λ using the EM (Expectation-Maximization) algorithm.

(a) (2 points) Given that each word w_i in document D_2 is generated independently from the mixture model, write down the log-likelihood of the whole document D_2. Is it easy to maximize this log-likelihood?

Solution: Let z and z_2 be a latent variable and is related by the equation: $\mathcal{P}(z=1) + \mathcal{P}(z=2) = 1$ The corresponding likelihood functions can be written as follows:

$$\mathcal{L}(\lambda|\theta,\theta_1) = \sum_{i=1}^{k} \log\left(\mathcal{P}(w_i|\theta_1).\mathcal{P}(z=1) + \mathcal{P}(w_i|\theta).\mathcal{P}(z=2)\right)$$

which can be written as:

$$\mathcal{L}(\lambda|\theta,\theta_1) = \sum_{i=1}^{k} \log\left(\mathcal{P}(w_i|\theta_1).(\lambda) + \mathcal{P}(w_i|\theta).(1-\lambda)\right)$$

As we have the sum of terms inside the log function, hence it is not easy to find global optima for the likelihood.

Differentiating likelihood withλ gives:

$$\frac{\mathcal{P}(w_i|\theta_i) - \mathcal{P}(w_i|\theta)}{\mathcal{P}(w_i|\theta_i).\lambda + \mathcal{P}(w_i|\theta).(1-\lambda)}$$

Alternatively we use EM algorithm for finding the local maxima. And starting from multiple points, we may hopefully reach global maxima. And it is not always guaranteed that we will find golbal maxima.

(b) (4 points) Write down the E-step and M-step updating formulas for estimating λ. Show your derivation of these formulas.

Solution:
E-Step: Getting individual prior probabilities of latent variables.

$$\mathcal{P}(z=0|w_i;\lambda) = \frac{\hat{\lambda}\mathcal{P}(w_i|\theta_1)}{\hat{\lambda}\mathcal{P}(w_i|\theta_1) + (1-\hat{\lambda})\mathcal{P}(w_i|\theta)}$$

$$\mathcal{P}(z=1|w_i;\lambda) = \frac{\hat{\lambda}\mathcal{P}(w_i|\theta_1)}{(\hat{\lambda})\mathcal{P}(w_i|\theta) + (1-\hat{\lambda})\mathcal{P}(w_i|\theta)}$$

M-Step: Maximising expectation for θ_t and the point at which we get maxima is set as θ_{t+1}.

$$\theta_{t+1} = \arg\max_{\theta}\left(\sum_{x in \mathcal{D}}\sum_{z}\mathcal{P}(z|x;\theta_t)\right)$$

$$\begin{aligned} g(\lambda) = \sum_{i-1}^{k}\left(\frac{\hat{\lambda}P\left(w_i \mid \theta_1\right)}{\hat{\lambda}P\left(w_i \mid \theta_1\right) + \left(1-\hat{\lambda}\right)P\left(w_i \mid \theta\right)} + \log(\lambda P\left(w_i \mid \hat{\theta}\right)\right) \\ + \left(\frac{(1-\lambda_t)\,P\left(w_i \mid \theta\right))}{\lambda_1 P\left(w_i \mid \theta_1\right) + (1-\lambda_1)\cdot P\left(w_i \mid \theta\right)} + \log\left((1-\lambda \mid P\left(w_i \mid \theta\right))\right)\right) \end{aligned} \tag{10}$$

In order to obtain the argmax of $g(\lambda$ we need to differentiate w.r.t λ. After differentiating we get,

$$\hat{\lambda}_{updated} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\hat{\lambda}\mathcal{P}(w_i|\theta_1)}{\hat{\lambda}\mathcal{P}(w_i|\theta_1) + (1-\hat{\lambda})\mathcal{P}(w_i|\theta)}\right)$$

(c) (4 points) In the previous parts of the question, we assume that the background language model $P(w|\theta)$ is known. How will your E-step and M-step change if you do not know the parameter θ and only know θ_1? Show your derivation.

Solution:

(d) (3 points) [BONUS] The previous parts of the question deal with MLE based density estimation. If you were to employ a Bayesian estimation method to infer λ, how will you proceed? That is, what prior would you choose for λ, and what is the the formula for the posterior? Is this posterior easily computable (i.e., has a closed-form expression or can be computed efficiently)? You can assume that both $P(w|\theta_1)$ and $P(w|\theta)$ are known and only λ is not known.

Solution: I would choose the appropriate model as the conjugate prior in a Bayesian setting. Doing so, we get the the posterior model in the same form as the prior, hence we get a closed form solution.
But it is not easy to always find a suitable conjugate prior for a given likelihood model.

5. (10 points) [DENSITY ESTIMATION - THE ONE RING TO RULE THEM ALL!] With density estimation ring already in your finger, you have all you need to master simple linear regression (even before seeing regression formally in class). Simple linear regression is a model that assumes a linear relationship between an input (aka independent) variable x and an output (aka dependent) variable y. Let us assume that the available set of observations, $\mathbb{D} = \{x_i, y_i\}_{i=1}^n$, are iid samples from the following model that captures the relationship between y and x:

$$y_i = w_0 + w_1 x_i + \epsilon_i; \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In this model, note that x_i is not a random variable, whereas ϵ_i and hence y_i are random variables, with ϵ_i being modeled as a Gaussian noise that is independent of each other and doesn't depend on x_i. Value of σ is assumed to be known for simplicity.
We would like to learn the parameters $\theta = \{w_0, w_1\}$ of the model, i.e., we would like to use MLE to estimate the exact parameter values or Bayesian methods to infer the (posterior) probability distribution over the parameter values.

(a) (2 points) Compute the probability distribution $P(y_i|x_i, \theta)$, and use it to write down the log likelihood of the model.

Solution: Given the distribution of $\epsilon \sim \mathcal{N}(0, \sigma^2)$,
For a given x_i and θ,

$$y_i = (w_0 + w_1 x_i) + \epsilon_i$$

we can write,

$$\begin{aligned} E[y_i|x_i,\theta] &= E[(w_o + w_1 x_i + \epsilon_i)|x_i,\theta] \\ &= E[w_o|x_i,\theta] + E[w_1 x_i|x_i,\theta] + E[\epsilon_i|x_i,\theta] \\ &= w_o + w_1 x_i + 0 = w_o + w_1 x_i \end{aligned}$$

$$\begin{aligned} V[y_i|x_i,\theta] &= V[(w_o + w_1 x_i + \epsilon_i)|x_i,\theta] \\ &= V[\epsilon_i|x_i,\theta] = \sigma^2 \end{aligned}$$

Now we know the mean and variance of $y_i|x_i,\theta$, we can write the form of distribution of $y_i|x_i,\theta$
$y_i|x_i,\theta \sim \mathcal{N}(w_o + w_1 x_i, \sigma^2)$

$$p(y_i|x_i,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i - w_o - w_1 x_i)^2}{2\sigma^2}}$$

Now log likelihood can be written as (assuming that y_i's are sampled from iid dist.):

$$-\log(\mathcal{L}(y_i; x_i, \theta)) = \sum_{i=1}^{n} \left(\log(\sigma\sqrt{2\pi}) + \frac{(y_i - w_o - w_1 x_i)^2}{2\sigma^2} \right) \tag{11}$$

(b) (3 points) Derive the ML estimates for w_0 and w_1 by optimizing the above log likelihood.

Solution: Differentiate expression 11 w.r.t w_o we get:

$$\sum_{i=1}^{n} 2(y_i - w_o - w_1 x_i) = 0$$

$$\sum_{i=1}^{n} (y_i) - n.\hat{w}_o - \hat{w}_1.(\sum_{i=1}^{n} (x_i)) = 0$$

$$\hat{w}_o = \frac{n.\bar{y} - \hat{w}_1.(n.\bar{x})}{n}$$

Differentiate expression 11 w.r.t w_1 we get:

$$\sum_{i=1}^{n} 2(y_i - w_o - w_1 x_i).x_i = 0$$

$$\sum_{i=1}^{n} \left[(y_i.x_i) - \hat{w}_o.x_i - \hat{w}_1.(x_i)^2)\right] = 0$$

$$\left[\overline{y.x} - \hat{w}_o.\bar{x} - n\hat{w}_1 \sum_{i=1}^{n}(x_i)^2)\right] = 0$$

$$\hat{w}_o = \frac{1}{\bar{x}}\left[\overline{xy} - \hat{w}_1\overline{x^2}\right]$$

Solving for w_1 we get:

$$\frac{1}{\bar{x}}\left[\overline{xy} - \hat{w}_1\overline{x^2}\right] = \bar{y} - \hat{w_1}\bar{x}$$

$$\overline{xy} - \hat{w}_1\overline{x^2} = \bar{x}.\bar{y} - \hat{w_1}\bar{x}^2$$

$$\hat{w_1} = \frac{\overline{xy} - \bar{x}.\bar{y}}{\overline{x^2} - \bar{x}^2}$$

And,

$$\hat{w_o} = \bar{y} - \bar{x}\left[\frac{\overline{xy} - \bar{x}.\bar{y}}{\overline{x^2} - \bar{x}^2}\right]$$

(c) (2 points) If σ is also not known before, derive the ML estimate for σ.

Solution: Differentiating the equation 11 with σ :

$$\sum_{i=1}^{n}\left[\frac{1}{\sigma} + (-2).\frac{(y_i - w_o - w_1.x_i)^2}{2\sigma^3}\right] = 0$$

$$\frac{n\sigma^2}{1} = \sum_{i=1}^{n}\left[(y_i - \hat{w}_o - \hat{w}_1.x_i)^2\right]$$

$$\sigma^2 = \frac{1}{n}.\sum_{i=1}^{n}\left[(y_i - \hat{w}_o - \hat{w}_1.x_i)^2\right]$$

(d) (3 points) For Bayesian inference, assume that the parameters w_0, w_1 are independent of each other and follow the distributions $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\mu_1, \sigma_1^2)$ respectively. Compute the posterior

distributions for each parameter. How does the mode of this posterior (i.e., MAP estimate) relate to the MLE of w_0 and w_1 derived above?

Solution: Posterior Probability of $\mathcal{P}(w_0|y_i)$ can be calculated as follows:

$$\mathcal{P}(w_0|y_i) = \frac{\mathcal{P}(y_i|w_0)\mathcal{P}(w_0)}{\mathcal{P}(w_0|y_i)}$$

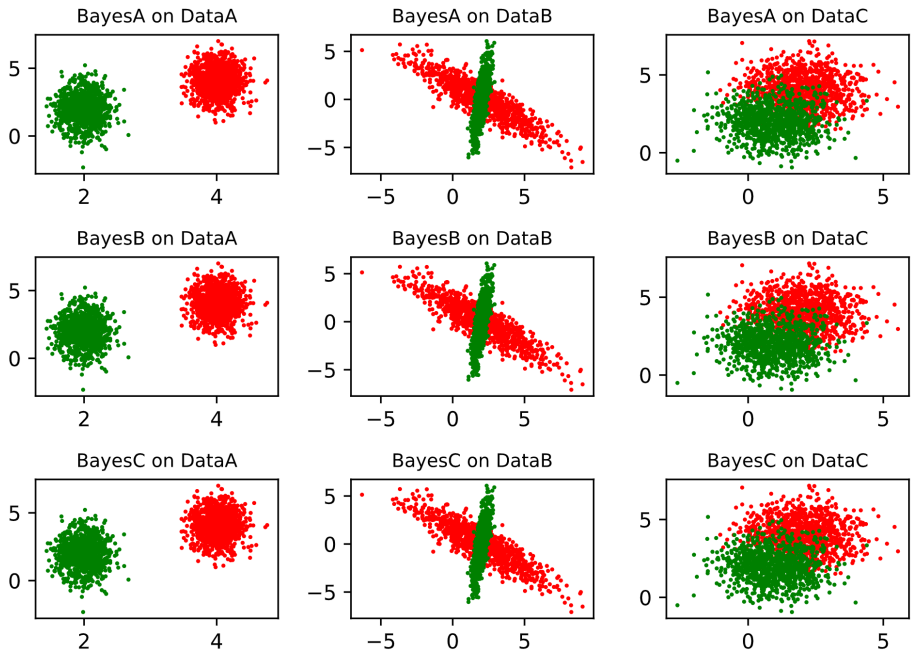
6. (10 points) [LET'S ROLL UP YOUR CODING SLEEVES...] **Learning Binary Bayes Classifiers from data via Density Estimation**
Derive Bayes classifiers under assumptions below and employing maximum likelihood approach to estimate class prior/conditional densities, and return the results on a test set.

   1. **BayesA** Assume $X|Y = -1 \sim \mathcal{N}(\mu_-, I)$ and $X|Y = 1 \sim \mathcal{N}(\mu_+, I)$
   2. **BayesB** Assume $X|Y = -1 \sim \mathcal{N}(\mu_-, \Sigma)$ and $X|Y = 1 \sim \mathcal{N}(\mu_+, \Sigma)$
   3. **BayesC** Assume $X|Y = -1 \sim \mathcal{N}(\mu_-, \Sigma_-)$ and $X|Y = 1 \sim \mathcal{N}(\mu_+, \Sigma_+)$

   Please see this folder for the template .ipynb file containing the helper functions, and you've to add the missing code to this file (specifically, three functions function_for_A, function_for_B and function_for_C, and associated plotting/ROC code snippets) to implement the above three algorithms for the three datasets given in the same folder.

   Please provide your results/answers in the pdf file you upload to GradeScope, but please submit your code separately in this moodle link. The code submitted should be a rollno.zip file containing two files: rollno.ipynb file (including your code as well as the exact same results/plots uploaded to Gradescope) and the associated rollno.py file.

   (a) (3 points) Plot all the classifiers (3 classification algorithms on 3 datasets = 9 plots) on a 2D plot, Add the training data points also on the plots. (Color the positively classified area light green, and negatively classified area light red as in Fig 4.5 in Bishop's book).

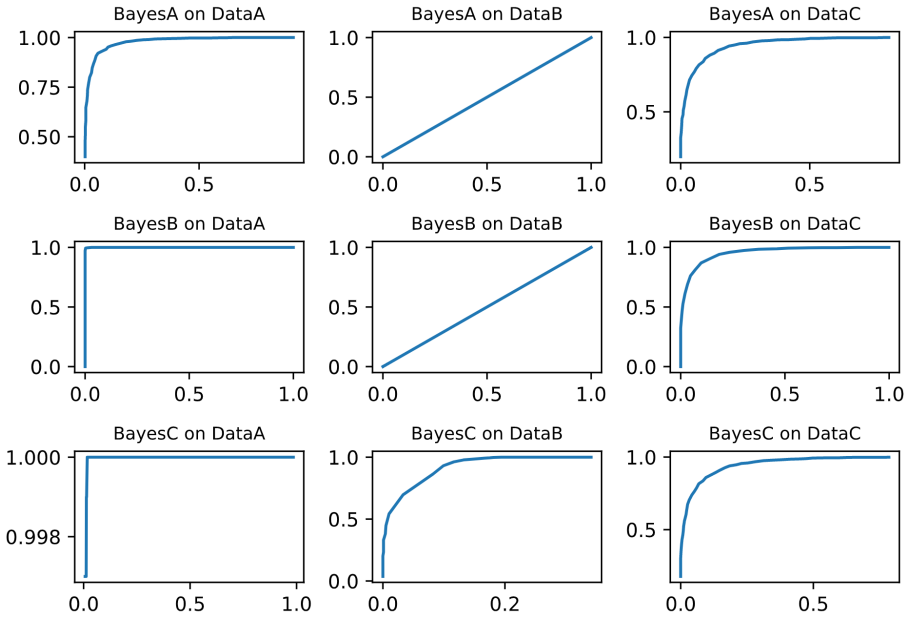


Solution:

Note: I didn't get time to complete the plotting of background which represents the classifier's prediction for all values of x. Hence, I'm mentioning the process here, we can do it by considering a meshgrid of formed by the training data space which is bounded by the min. and max. of the training data (X)

(b) (3 points) Give the ROC curves for all the classifiers. Note that a ROC curve plots the FPR (False Positive Rate) on the x-axis and TPR (True Positive Rate) on the y-axis. (9 plots)

Solution:



(c) (2 points) Provide the error rates for the above classifiers (three classifiers on the three datasets as 3×3 table, with appropriately named rows and columns).

Solution: Below is the table for error rates, calculated when the different classifiers are trained, tested on different Data sets.

Error Rate in (%)			
Error Rate	**BayesA**	**BayesB**	**BayesC**
DatasetA	9.80	22.9	22.55
DatasetB	50.85	50.40	7.45
DatasetC	11.75	11.65	11.79

(d) (2 points) Summarise and explain your observations based on your plots and the assumptions given in the problem. Also briefly comment whether a non-parametric density estimation approach could have been used to solve this problem, and if so, what the associated pros/cons are compared to the parametric MLE based approach you have implemented.

Solution: We get higher error rate while operating on datasetB. This is because the training data of the two classifiers is coinciding. Hence there is a inherent error from the data.

1. DatasetA is the easiest to operate with, as the data points ("clusters") are far away and the decision boundary is easiest to deduce. 2. Among the three classifiers, BayesC will make the accurate decision boundary given the training data set. But in the above case BayesA is performing better on datasets. This might be a co-incidence as the testing data might be

'favourable' to the classifier - BayesA.

We might not get good results with non-parametric density estimation, especially in case of when operated on datasetB as both type of data are clustered in close to each other.