IITM-CS5691 : Pattern Recognition and Machine Learning
Assignment 1

Release Date: March 9, 2021
Due Date : March 19, 23:59

Roll No: EE18B001
Name: Abishek S
Collaborators (if any):
References (if any): Class Notes, Bishop

- 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:
   Let us define the events,
   F : event that a fair coin is picked
   $\overline{\mathbf{F}}$: event that the unfair coin is picked
   E : event that heads has occurred 10 times in a row
   A : event that the next toss of the same coin lands heads

By Total Probability law,

$$\begin{aligned} P(A \mid E) &= P(A \mid F, E)P(F \mid E) + P(A \mid \bar{F}, E)P(\bar{F} \mid E) \\ &= P(A \mid F, E)P(F \mid E) + P(A \mid \bar{F}, E)(1 - P(F \mid E)) \end{aligned} \tag{1}$$

By Bayes theorem,

$$\begin{aligned} P(F \mid E) &= \frac{P(E \mid F)P(F)}{P(E \mid F)P(F) + P(E \mid \bar{F})P(\bar{F})} \\ &= \frac{(0.5)^{10}(0.999)}{(0.5)^{10}(0.999) + (1)^{10}(0.001)} = 0.49382 \end{aligned} \tag{2}$$

Clearly, events A and E are conditionally independent on event F, (i.e)

$$\begin{aligned} P(A \mid F, E) &= P(A \mid F) = 0.5 \\ P(A \mid \bar{F}, E) &= P(A \mid \bar{F}) = 1 \end{aligned} \tag{3}$$

Substituting 2 and 3 in 1,

$$P(A \mid E) = (0.5)P(F \mid E) + (1)(1 - P(F \mid E)) = 0.753$$

We can call $P(F \mid E)$ as a **posterior probability** because it is the probability of a "posterior" event (that happened earlier) computed from it's effect using Bayes theorem. $P(A \mid E)$ is then almost like a predictive distribution where we marginalize over all possible causes that could have happened to predict the possibility of a future event.

(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:
Proving that the ML estimates correspond to global maxima of the likelihood function :
Since the sample are i.i.d,

$$\begin{aligned} P(\mathbf{X} \mid \mu, \sigma^2) &= \prod_{i=1}^{n} P(x_i \mid \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \mathcal{N}(\mu, \sigma^2) \end{aligned}$$

and the log likelihood is given as :

$$\log P(\mathbf{X} \mid \mu, \sigma^2) = -\frac{n}{2}\log 2\pi - n\log\sigma - \sum_{i=1}^{n}\frac{1}{2\sigma^2}(x_i - \mu)^2 \tag{1}$$

The solution of the maximum likelihood maximizes the log likelihood, or in other words, ML estimates of μ and σ are obtained by setting the gradient of log likelihood to zero as :

$$\begin{aligned}
&\frac{\partial}{\partial\mu}\log P(\mathbf{X} \mid \mu, \sigma^2) = 0 \\
&\implies \frac{\partial}{\partial\mu}\sum_{i=1}^{n}\frac{1}{2\sigma^2}(x_i - \mu)^2 = 0 \\
&\implies \sum_{i=1}^{n}(x_i - \mu_{ML}) = 0 \\
&\implies \mu_{ML} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}
\end{aligned} \tag{2}$$

$$\begin{aligned}
&\frac{\partial}{\partial\sigma}\log P(\mathbf{X} \mid \mu, \sigma^2) = 0 \\
&\implies \frac{\partial}{\partial\sigma}(-n\log\sigma - \sum_{i=1}^{n}\frac{1}{2\sigma^2}(x_i - \mu)^2) = 0 \\
&\implies -n/\sigma_{ML} + \sum_{i=1}^{n}\frac{1}{\sigma_{ML}^3}(x_i - \mu_{ML})^2 = 0 \qquad (3) \\
&\implies \sigma_{ML}^2 = \frac{\sum_{i=1}^{n}(x_i - \mu_{ML})^2}{n} \qquad (4)
\end{aligned}$$

Since the log likelihood is a convex function in μ and σ, and there exists only one solution for μ_{ML} given by 1 and σ_{ML} is also uniquely defined for a given μ_{ML} as evident from 2, the solution corresponds to the global maxima.
We can justify that the solution corresponds to global maxima in another way. The only term in 1 that μ depends on is the third term. For an arbitrary (fixed) σ, to maximise the log likelihood, the third term needs to be maximised. In other words,

$$\sum_{i=1}^{n}(x_i - \mu)^2$$

needs to be minimised (since σ is fixed, it can be ignored in the third term while varying

μ). We can observe the following :

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\mu)^2 \\
&= \sum_{i=1}^{n}[(x_i-\overline{x})^2+(\overline{x}-\mu)^2+2(x_i-\overline{x})(\overline{x}-\mu)] \\
&= \sum_{i=1}^{n}[(x_i-\overline{x})^2+(\overline{x}-\mu)^2]+2(\overline{x}-\mu)\sum_{i=1}^{n}(x_i-\overline{x}) \qquad (5) \\
&= \sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(\overline{x}-\mu)^2 \\
&\geqslant \sum_{i=1}^{n}(x_i-\mu_{ML})^2 \qquad (6)
\end{aligned}$$

(where in 4 the third term vanishes because $n\mu_{ML} = n\overline{x} = \sum_{i=1}^{n} x_i$).
Since the value of μ corresponding to global maxima does not depend on σ, we can maximise the log likelihood by varying σ and fixing μ as μ_{ML}. For maximising 1, considering only the terms containing σ, we need to minimise :

$$n\log\sigma+\frac{A}{2\sigma^2} \qquad (7)$$

(where $A = \sum_{i=1}^{n}(x_i-\mu)^2$, and $A >= 0$ obviously).
The extremum points for this is obtained by differentiating the expression once wrt σ and equating to zero, as in 4. But there is another solution that yields zero derivative, which is $\sigma = \infty$ and this can be verified by substituting it in 3.
But when $\sigma = \infty$, 7 becomes : $n\log\infty + 0 = \infty$, whereas the σ_{ML} solution yields a finite value and hence clearly is minimum ($<= \infty$). Hence the Maximum Likelihood solutions correspond to the global maxima of the log likelihood function, and since log is a monotonically increasing function, the ML estimates correspond to the global maxima of the likelihood function.

Deriving bias of the MLE of μ and σ^2 :
The μ_{ML} and σ_{ML} obtained are functions of the i.i.d data $\mathbf{X} = x_{i\,i=1}^{n}$, we take the expectation

of the values over the data first.

$$\begin{aligned}
\mathbb{E}_{x_1,x_2,\ldots,x_n}[\mu_{ML}] = \mathbb{E}_{\mathbf{x}}[\mu_{ML}] &= \mathbb{E}_{\mathbf{x}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}}[x_i] && \text{(by linearity of expectations)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu && (x_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)\ ) \\
&= \mu && (8)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}[\sigma^2_{ML}] &= \mathbb{E}_{\mathbf{x}}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{ML})^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}}[x_i^2 + \mu^2_{ML} - 2x_i\mu_{ML}] && \text{(by linearity of expectations)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_{\mathbf{x}}[x_i^2] + \mathbb{E}_{\mathbf{x}}[\mu^2_{ML}] - 2\mathbb{E}_{\mathbf{x}}\left[x_i\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right]\right) && (9)
\end{aligned}$$

Now, let us evaluate the terms of 9 individually.

$$\mathbb{E}_{\mathbf{x}}[x_i^2] = var(x) + (\mathbb{E}_{\mathbf{x}}[x_i])^2 = \sigma^2 + \mu^2 \tag{10}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}[\mu^2_{ML}] &= \mathbb{E}_{\mathbf{x}}[(\frac{1}{n}\sum_{i=1}^{n} x_i)^2] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}_{\mathbf{x}}[x_i x_j] && \text{(by linearity of expectations)} \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathbb{E}_{\mathbf{x}}[x_i x_j] + \sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}}[x_i^2]\right) && \text{(by linearity again)} \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathbb{E}_{\mathbf{x}}[x_i]\mathbb{E}_{\mathbf{x}}[x_j] + \sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}}[x_i^2]\right) && (x_i, x_j \text{ - independent samples}) \\
&= \frac{1}{n^2}\big(n(n-1)\mu^2 + n(\mu^2 + \sigma^2)\big) && (x_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \text{ and from 9}) \\
&= \mu^2 + \frac{\sigma^2}{n} && (11)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}\left[x_i\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right] &= \mathbb{E}_{\mathbf{X}}\left[\frac{1}{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} x_i x_j + x_i^2\right)\right] \\
&= \frac{1}{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} \mathbb{E}_{\mathbf{X}}[x_i x_j] + \mathbb{E}_{\mathbf{X}}[x_i^2]\right) && \text{(by linearity of expectations)} \\
&= \frac{1}{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} \mathbb{E}_{\mathbf{X}}[x_i]\mathbb{E}_{\mathbf{X}}[x_j] + \mathbb{E}_{\mathbf{X}}[x_i^2]\right) && (x_i, x_j \text{ - independent samples}) \\
&= \frac{1}{n}\left((n-1)\mu^2 + (\mu^2+\sigma^2)\right) && (x_i \overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2)\ ) \\
&= \mu^2 + \frac{\sigma^2}{n} && (12)
\end{aligned}$$

Substituting 10, 11 and 12 in 9,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\sigma^2_{ML}] &= \frac{1}{n}\sum_{i=1}^{n}\left((\sigma^2+\mu^2) + \left(\mu^2+\frac{\sigma^2}{n}\right) - 2\left(\mu^2+\frac{\sigma^2}{n}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left((\sigma^2+\mu^2) - \left(\mu^2+\frac{\sigma^2}{n}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{(n-1)}{n}\sigma^2 \\
&= \frac{(n-1)}{n}\sigma^2 && (13)
\end{aligned}$$

Bias of an estimator of a parameter $\hat{Y}$ (here) is $\mathbb{E}_{\mathbf{X}}[\hat{Y}] - Y$, where Y is the true value of the parameter.
Hence, bias in MLE mean $(\mu_{ML}) = \mathbb{E}_{\mathbf{X}}[\mu_{ML}] - \mu = \mu - \mu = 0$ (using result from 8).
Bias in MLE variance $(\sigma^2_{ML}) = \mathbb{E}_{\mathbf{X}}[\sigma^2_{ML}] - \sigma^2 = \frac{(n-1)}{n}\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n}$ (using result from 13).

(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:
Proving $\mathbb{H}$ is a subspace of $\mathbb{R}^d$:
To prove a given set is a subspace, two conditions need to be satisfied :

1. Presence of additive identity (i.e) 0 element.

- Hyperplane $\mathbb{H}$ passes through zero, hence additive identity zero $\in \mathbb{H}$.

2. Closure under linear combination (i.e) $u, v \in \mathbb{H} \implies \alpha_1 u + \alpha_2 v \in \mathbb{H} \quad \forall\ \alpha_1, \alpha_2 \in \mathbb{R}$ (since we are dealing with real numbers, the scalars are real).

   - Since $\mathbb{H}$ is a hyperplane through zero, any vector on it satisfies the equation $w^T x = 0$, where $w \in \mathbb{R}^d$ and $w \neq 0$.
   - Suppose $u, v \in \mathbb{H} \implies w^T u = 0$ and $w^T v = 0$
   - The above implies $\alpha_1 w^T u = 0$ and $\alpha_2 w^T v = 0$ for arbitrary α_1 and $\alpha_2 \in \mathbb{R}$. This implies $\alpha_1 w^T u + \alpha_2 w^T v = 0$
   - $\implies w^T(\alpha_1 u) + w^T(\alpha_2 v) = 0$ (scalar multiplication is commutative)
   - $\implies w^T(\alpha_1 u + \alpha_2 v) = 0$ (distributive property of vector multiplication over addition)
   - The above statement implies $\alpha_1 u + \alpha_2 v \in \mathbb{H}$.

Hence $\mathbb{H}$ is a subspace of $\mathbb{R}^d$.

Proving $\dim(\mathbb{H}) = d - 1$ **:**
Let the hyperplane equation be given by $w^T x = 0$, where $w \in \mathbb{R}^d$ and $w \neq 0$.
Define the matrix $W = w^T = [w_1\ w_2\ \cdots\ w_n]$.
Matrix W can be thought of as a linear map from $\mathbb{R}^d \longmapsto \mathbb{R}$.
Clearly $\mathbb{H}$ is then the set $\{x \in \mathbb{R}^d \mid Wx = 0\}$, and hence $\dim(\mathbb{H}) = \text{nullity}(W)$.
Since not all w_i's are zero, $\text{rank}(W) = 1$ and $\dim(\mathbb{R}^d) = d$.
By the fundamental theorem of linear algebra (also call the **rank-nullity theorem**),

$$\text{nullity}(W) = \dim(\mathbb{R}^d) - \text{rank}(W) = d - 1$$

(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:
The likelihood function of the Gaussian mixture model (GMM) is **not a convex function** wrt to it's parameters. Let us prove this by showing a counter-example.
Since π_1, π_2 are the mixing coefficients, $\pi_1 + \pi_2 = 1$.
Let there be N i.i.d samples given; **x** denotes $\{x_i\}_{i=1}^N$.
The likelihood function is then given as function of the parameters (jointly denoted by

$\theta = [\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2])$ as :

$$\mathcal{L}(\theta; \mathbf{x}) = \mathcal{L}(\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2; \mathbf{x}) = \prod_{i=1}^{N} \Big(\pi_1 \mathcal{N}(x \mid \mu_1, \sigma_1^2) + (1 - \pi_1) \mathcal{N}(x \mid \mu_2, \sigma_2^2) \Big)$$

where $0 \leqslant \pi_1 \leqslant 1;\ \sigma_1, \sigma_2 \geqslant 0;\ \mu_1, \mu_2 \in \mathbb{R}$

The test for convexity for a function f is :

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y) \qquad \forall\, x, y \in domain(f) \tag{1}$$

Define $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$. Let us choose

- $\lambda = \frac{1}{2}$
- $\theta_1 = [\pi_1 = \pi,\ \mu_1 = \overline{x},\ \mu_2 = \overline{x},\ \sigma_1 = \sigma,\ \sigma_2 = \sigma]$
- $\theta_2 = [\pi_2 = \pi,\ \mu_1 = \overline{x} + \frac{2k\sigma}{\sqrt{N}},\ \mu_2 = \overline{x} + \frac{2k\sigma}{\sqrt{N}},\ \sigma_1 = \sigma,\ \sigma_2 = \sigma]$

for arbitrary constants $\pi \in [0, 1];\ \sigma \in [0, \infty);\ k \in \mathbb{R}$ and given N datapoints ${x_i}_{i=1}^{N} \in \mathbb{R}$.
Let us perform convexity test for function $f = \mathcal{L}(\theta; x)$.
(For simplicity we denote it as $\mathcal{L}(\theta)$)

LHS :

$$
\begin{aligned}
\mathcal{L}(\lambda\theta_1 + (1-\lambda)\theta_2) &= \mathcal{L}\left(\frac{\theta_1+\theta_2}{2}\right) \\
&= \prod_{i=1}^{N}\left(\frac{\pi+\pi}{2}\right)\mathcal{N}\left(\left(\frac{\overline{x}+\overline{x}+\frac{2k\sigma}{\sqrt{N}}}{2}\right),\left(\frac{\sigma+\sigma}{2}\right)^2\right) \\
&\quad + \left(1-\frac{\pi+\pi}{2}\right)\mathcal{N}\left(\left(\frac{\overline{x}+\overline{x}+\frac{2k\sigma}{\sqrt{N}}}{2}\right),\left(\frac{\sigma+\sigma}{2}\right)^2\right) \\
&= \prod_{i=1}^{N}\pi\,\mathcal{N}\left(\overline{x}+\frac{k\sigma}{\sqrt{N}},\sigma^2\right) + (1-\pi)\,\mathcal{N}\left(\overline{x}+\frac{k\sigma}{\sqrt{N}},\sigma^2\right) \\
&= \prod_{i=1}^{N}\mathcal{N}\left(\overline{x}+\frac{k\sigma}{\sqrt{N}},\sigma^2\right) \\
&= \prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(x-\left(\overline{x}+\frac{k\sigma}{\sqrt{N}}\right)\right)^2}{2\sigma^2}\right) \\
&= \prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2}+\frac{(x-\overline{x})\frac{k}{\sqrt{N}}}{\sigma}-\frac{k^2}{2N}\right) \\
&= \frac{\exp\left(-\frac{N(x-\overline{x})^2}{2\sigma^2}\right)}{\left(\sigma\sqrt{2\pi}\right)^N}\exp\left(-\frac{k^2}{2}\right) \qquad (2)
\end{aligned}
$$

where in the last step, the cross term $\frac{(x-\overline{x})\frac{k}{\sqrt{N}}}{\sigma}$ in the exponential vanishes when product is taken because of the definition of $\overline{x}$.

RHS :

$$
\begin{aligned}
\lambda\mathcal{L}(\theta_1) + (1-\lambda)\mathcal{L}(\theta_2) &= \frac{1}{2}\big(\mathcal{L}(\theta_1) + \mathcal{L}(\theta_2)\big) \\
&= \prod_{i=1}^{N}\left(\frac{\pi}{2}\,\mathcal{N}(\overline{x}, \sigma^2) + \frac{(1-\pi)}{2}\,\mathcal{N}(\overline{x}, \sigma^2)\right) \\
&\quad + \prod_{i=1}^{N}\left(\frac{\pi}{2}\,\mathcal{N}\left(\overline{x} + \frac{2k\sigma}{\sqrt{N}}, \sigma^2\right) + \frac{(1-\pi)}{2}\,\mathcal{N}\left(\overline{x} + \frac{2k\sigma}{\sqrt{N}}, \sigma^2\right)\right) \\
&= \frac{1}{2}\prod_{i=1}^{N}\left(\mathcal{N}(\overline{x}, \sigma^2)\right) + \frac{1}{2}\prod_{i=1}^{N}\left(\mathcal{N}\left(\overline{x} + \frac{2k\sigma}{\sqrt{N}}, \sigma^2\right)\right) \\
&= \frac{1}{2\big(\sigma\sqrt{2\pi}\big)^N}\prod_{i=1}^{N}\left(\exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2}\right)\right) \\
&\quad + \frac{1}{2\big(\sigma\sqrt{2\pi}\big)^N}\prod_{i=1}^{N}\left(\exp\left(-\frac{\left(x - \left(\overline{x} + \frac{2k\sigma}{\sqrt{N}}\right)\right)^2}{2\sigma^2}\right)\right) \\
&= \frac{\exp\left(-N\frac{(x-\overline{x})^2}{2\sigma^2}\right)}{2\big(\sigma\sqrt{2\pi}\big)^N} \\
&\quad + \frac{1}{2\big(\sigma\sqrt{2\pi}\big)^N}\prod_{i=1}^{N}\left(\exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2} + \frac{(x-\overline{x})\frac{2k}{\sqrt{N}}}{\sigma} - \frac{2k^2}{N}\right)\right) \\
&= \frac{\exp\left(-\frac{N\,(x-\overline{x})^2}{2\sigma^2}\right)}{2\big(\sigma\sqrt{2\pi}\big)^N} + \frac{\exp\left(-\frac{N\,(x-\overline{x})^2}{2\sigma^2}\right)}{2\big(\sigma\sqrt{2\pi}\big)^N}\exp\left(-2k^2\right) \\
&= \frac{\exp\left(-\frac{N\,(x-\overline{x})^2}{2\sigma^2}\right)}{2\big(\sigma\sqrt{2\pi}\big)^N}\left(1 + \exp\left(-2k^2\right)\right) \qquad (3)
\end{aligned}
$$

In RHS also the cross term in the exponential vanishes when taking product due to the way $\overline{x}$ is defined.
Using a graphic calculator I solved the inequality **LHS** $>$ **RHS** for k and bactracked to obtain the counter-example.
For $k = 0.7$ (k can be chosen to be any real value between 0 and 1.1, but for example choosing a particular value here for illustrating),

LHS :

$$\frac{\exp\left(-\frac{N(x-\bar{x})^2}{2\sigma^2}\right)}{\left(\sigma\sqrt{2\pi}\right)^N}\exp\left(-\frac{k^2}{2}\right) = \frac{\exp\left(-\frac{N(x-\bar{x})^2}{2\sigma^2}\right)}{\left(\sigma\sqrt{2\pi}\right)^N}\exp\left(-\frac{(0.7)^2}{2}\right)$$
$$= \frac{\exp\left(-\frac{N(x-\bar{x})^2}{2\sigma^2}\right)}{\left(\sigma\sqrt{2\pi}\right)^N}(0.7827) \tag{4}$$

RHS :

$$\frac{\exp\left(-\frac{N(x-\bar{x})^2}{2\sigma^2}\right)}{2\left(\sigma\sqrt{2\pi}\right)^N}\left(1+\exp\left(-2k^2\right)\right) = \frac{\exp\left(-\frac{N(x-\bar{x})^2}{2\sigma^2}\right)}{2\left(\sigma\sqrt{2\pi}\right)^N}\left(1+\exp\left(-2(0.7)^2\right)\right)$$
$$= \frac{\exp\left(-\frac{N(x-\bar{x})^2}{2\sigma^2}\right)}{2\left(\sigma\sqrt{2\pi}\right)^N}(1+0.3753)$$
$$= \frac{\exp\left(-\frac{N(x-\bar{x})^2}{2\sigma^2}\right)}{\left(\sigma\sqrt{2\pi}\right)^N}(0.6876) \tag{5}$$

From 4 and 5 clearly, LHS $>$ RHS and hence equation 1 is not satisfied, indicating $\mathcal{L}(\theta; \mathbf{x})$ is not a convex function.

(e) (2 points) Show that there always exists a solution for the system of equations, $A^TAx = A^Tb$, 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:
Proving existence of solution for $A^TAx = A^Tb$:
Observations about the matrix A^TA and resulting conclusions :

- It is a square matrix $\in \mathbb{R}^{m\times m}$.
- $(A^TA)^T = A^TA$, hence it is real and symmetric $\implies$ it is a self-adjoint ($\implies$ $normal$) operator.

We know that $nullity(BA) = nullity(A) + dim(range(A) \cap null(B))$ and if $range(A) \cap null(B) = 0$, $null(BA) = null(A)$.
So, $range(A)^\perp = null(A^T) \implies null(A^TA) = null(A)$.

$$range(A^TA) = null(A^TA)^\perp = null(A)^\perp = range(A^T)$$

$$A^T b \in range(A^T) \implies A^T b \in range(A^T A)$$

Hence, $A^T b$ can be written as $c_1 (A^T A)_{\cdot,1} + c_2 (A^T A)_{\cdot,2} + \cdots + c_m (A^T A)_{\cdot,m}$

$$\text{since} \quad columnspace(A^T A) = span \left\{ \begin{pmatrix} | \\ (A^T A)_{\cdot,1} \\ | \end{pmatrix} \quad , \quad \cdots \quad , \quad \begin{pmatrix} | \\ (A^T A)_{\cdot,m} \\ | \end{pmatrix} \right\} \tag{2}$$

$(A^T A)_{\cdot,i}$ represents i^{th} column of the $m \times m$ matrix $A^T A$. The c_i's need not be unique as the column vectors of $A^T A$ need not be linearly independent.
If $c = [c_1, c_2, \cdots, c_m]^T$, clearly the equation 2

$$\implies A^T b = (A^T A)c \implies x = c \text{ is a solution to } A^T A x = A^T b$$

Hence there always exists a solution to $A^T A x = A^T b$ even though it may not be unique.

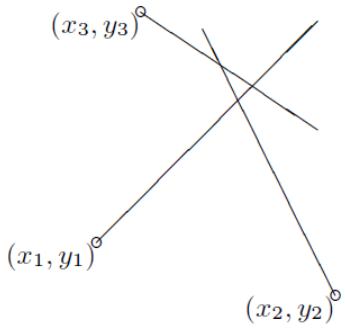
Showing x^* is the projection of b onto the column space (range) of A :
Consider a solution x^* to the equation $A^T A x = A^T b$.

$$\begin{aligned} & A^T A x^* = A^T b \implies A^T (A x^* - b) = 0 \\ & \implies (A x^* - b) \in nullspace(A^T) \\ & \implies (A x^* - b) \in range(A)^\perp \\ & \text{Now, } A x^* \in range(A) \implies b = A x^* + (A x^* - b) \end{aligned} \tag{2}$$

where $Ax^* \in range(A)$ and $(Ax^* - b) \in range(A)^\perp$.
Since $range(A)$ and $range(A)^\perp$ form a direct sum of the co-domain space of linear map represented by matrix A, b is uniquely written as combination of vectors in $range(A)$ and $range(A)^\perp$ as given in 2.
Hence Ax^* is clearly the projection of b on to the range (or column space) of A.

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:
(I assume the angles are in radians, to make computations easier and understandable, and handle inverse trigonometric functions easily. Angle in degrees can be obtained from that in radians simply by a multiplication factor of $\frac{180}{\pi}$).
$\tilde{\theta}_i = \theta_1 + \mathcal{N}(0, \sigma^2) \implies \tilde{\theta}_i \sim \mathcal{N}(\theta_i, \sigma^2) \quad \forall\, i \in \{1, 2, 3\}$.
Let us assume positive y-axis as the true north direction. Without loss of generality (and in accordance with the diagram shown), let (x_1, y_1) be in third quadrant, (x_2, y_2) in fourth quadrant and (x_3, y_3) in second quadrant all wrt (x, y). The true bearings (measured clockwise) can then be given from the coordinates by :

$$\theta_1 = \alpha_1 - \tan^{-1}\left(\frac{y_1 - y}{x_1 - x}\right) = \frac{3\pi}{2} - \tan^{-1}\left(\frac{y_1 - y}{x_1 - x}\right)$$
$$\theta_2 = \alpha_2 - \tan^{-1}\left(\frac{y_2 - y}{x_2 - x}\right) = \frac{\pi}{2} - \tan^{-1}\left(\frac{y_2 - y}{x_2 - x}\right)$$
$$\theta_3 = \alpha_3 - \tan^{-1}\left(\frac{y_3 - y}{x_3 - x}\right) = \frac{3\pi}{2} - \tan^{-1}\left(\frac{y_3 - y}{x_3 - x}\right)$$

The quadrant each point belongs to can be gauged from the measured $\tilde{\theta}_i$'s, as it deviates from true angle by a small error only (unless the point is very close to the axes). This will determine the offsets $\alpha_1, \alpha_2, \alpha_3$. For our assumption about quadrants of (x_i, y_i)'s we get $\alpha_1 = \frac{3\pi}{2}, \alpha_2 = \frac{\pi}{2}, \alpha_3 = \frac{3\pi}{2}$. The α_i's are arbitrary constants.
Each measurement is taken independently from it's distribution, so their join probability will

be product of each probability. Considering this fact, the likelihood function will be :

$$\begin{aligned} P(\{\tilde{\theta}_n\}_{n=1}^3 \mid x, y, \{x_n, y_n\}_{n=1}^3, \sigma) &= \prod_{i=1}^{n} P(\tilde{\theta}_i \mid x, y, x_i, y_i, \sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(\tilde{\theta}_i - \theta_i)^2}{2\sigma^2} \right) \end{aligned}$$

The log likelihood is then given by :

$$\begin{aligned} \mathcal{L}(x, y; \{\tilde{\theta}_n\}_{n=1}^3, \{x_n, y_n\}_{n=1}^3, \sigma) &= \log P(\{\tilde{\theta}_n\}_{n=1}^3 \mid x, y, \{x_n, y_n\}_{n=1}^3, \sigma) \\ &= -\frac{3}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\left(\tilde{\theta}_1 - \theta_1\right)^2 + \left(\tilde{\theta}_2 - \theta_2\right)^2 + \left(\tilde{\theta}_3 - \theta_3\right)^2 \right) \\ &= -\frac{3}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\left(\tilde{\theta}_1 - \alpha_1 + \tan^{-1}\left(\frac{y_1 - y}{x_1 - x}\right)\right)^2 \right. \\ &\quad + \left(\tilde{\theta}_2 - \alpha_2 + \tan^{-1}\left(\frac{y_2 - y}{x_2 - x}\right)\right)^2 \\ &\quad \left. + \left(\tilde{\theta}_3 - \alpha_3 + \tan^{-1}\left(\frac{y_3 - y}{x_3 - x}\right)\right)^2 \right) \end{aligned}$$

Maximum likelihood estimates are given by :

$$\begin{aligned} \{x_{ML}, y_{ML}\} &= \arg\max_{x,y} \ \mathcal{L}(x, y; \{\tilde{\theta}_n\}_{n=1}^3, \{x_n, y_n\}_{n=1}^3, \sigma) \\ &= \arg\max_{x,y} \ -\frac{3}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\left(\tilde{\theta}_1 - \alpha_1 + \tan^{-1}\left(\frac{y_1 - y}{x_1 - x}\right)\right)^2 \right. \\ &\quad \left. + \left(\tilde{\theta}_2 - \alpha_2 + \tan^{-1}\left(\frac{y_2 - y}{x_2 - x}\right)\right)^2 + \left(\tilde{\theta}_3 - \alpha_3 + \tan^{-1}\left(\frac{y_3 - y}{x_3 - x}\right)\right)^2 \right) \end{aligned}$$

We can find the ML estimates by setting the partial derivatives of the log likelihood wrt x and y to zero.

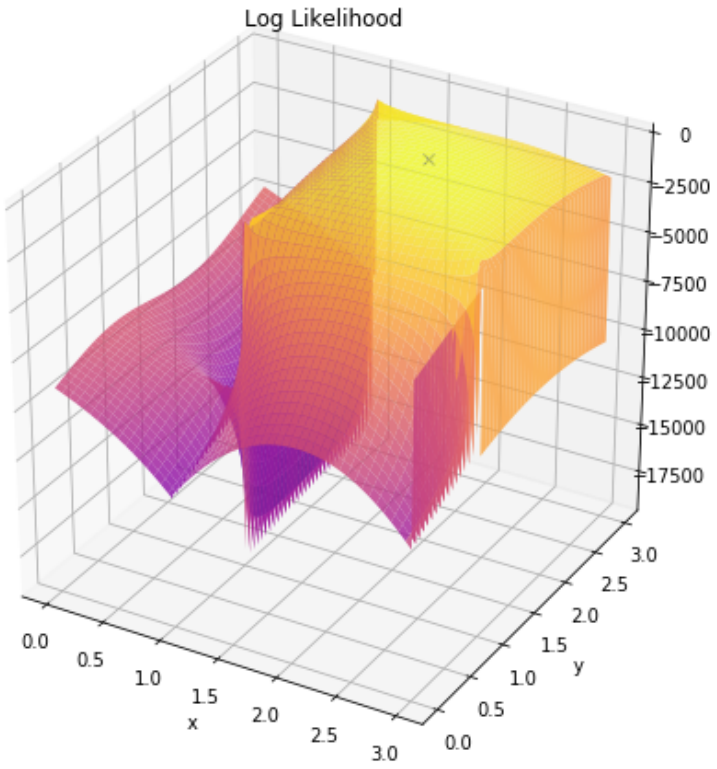
Setting partial derivative of log likelihood wrt x to zero :

$$\begin{aligned}
&\frac{\partial}{\partial x}\mathcal{L}(x, y; \{\tilde{\theta}_n\}_{n=1}^3, \{x_n, y_n\}_{n=1}^3, \sigma) = 0 \\
&\implies 2\left(\tilde{\theta}_1 - \alpha_1 + \tan^{-1}\left(\frac{y_1 - y}{x_1 - x}\right)\right) \cdot \frac{\partial}{\partial x}\tan^{-1}\left(\frac{y_1 - y}{x_1 - x}\right) \\
&\quad + 2\left(\tilde{\theta}_2 - \alpha_2 + \tan^{-1}\left(\frac{y_2 - y}{x_2 - x}\right)\right) \cdot \frac{\partial}{\partial x}\tan^{-1}\left(\frac{y_2 - y}{x_2 - x}\right) \\
&\quad + 2\left(\tilde{\theta}_3 - \alpha_3 + \tan^{-1}\left(\frac{y_3 - y}{x_3 - x}\right)\right) \cdot \frac{\partial}{\partial x}\tan^{-1}\left(\frac{y_3 - y}{x_3 - x}\right) = 0 \qquad \text{(by chain rule)} \\
&\implies \sum_{i=1}^{3}\left(\tilde{\theta}_i - \alpha_i + \tan^{-1}\left(\frac{y_i - y}{x_i - x}\right)\right) \cdot \left(\frac{y_i - y}{(x_i - x)^2 + (y_i - y)^2}\right) = 0
\end{aligned}$$

Similarly we can do wrt y. But,the function contains inverse trigonometric and rational polynomial terms in x and y, and clearly is complex, making it not easy to find closed form expressions for x and y. Hence we plot the log likelihood function assuming a set of values for $\{\tilde{\theta}_n\}_{n=1}^3, \{x_n, y_n\}_{n=1}^3, \sigma$ and show the nature of function.

Values chosen (all angles are in radians) :

- True $x = 2, y = 2$.
- $x_1 = 1, y_1 = 1; x_2 = 3, y_2 = 1; x_3 = 1, y_3 = 3$.
- $\implies \theta_1 = \frac{5\pi}{4}, \theta_2 = \frac{3\pi}{4}, \theta_3 = \frac{7\pi}{4}$.
- $\sigma = 1^\circ = \frac{\pi}{180}$.
- $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$ are randomly sampled from their respective normal distributions (given in the first line of solution) and for the plot below, the values are $3.92397716, 2.37988289$ and 5.52125742 respectively.
- Here, (x_1, y_1) is in third quadrant, (x_2, y_2) in fourth quadrant and (x_3, y_3) in second quadrant wrt (x, y). Hence, $\alpha_1 = \frac{3\pi}{2}, \alpha_2 = \frac{\pi}{2}, \alpha_3 = \frac{3\pi}{2}$.



We can see the blue cross denoting the true value $x = 2, y = 2$ is the point around which the log likelihood clearly peaks, asserting how close (if not equal) to the true solution, ML estimate arrives at.

We can see that the plot isn't purely concave or convex and is complex in nature. However, notice that in the neighbourhood of the true $(x, y) = (2, 2)$, the function is smooth (and even seems to be purely concave). Hence only the values of x and y that make partial derivatives go to zero maximises the log likelihood. This value is given by the ML estimate (since the function seems purely concave near the true (x, y), there will be a unique maxima). Any other solution can only be as good as the ML estimate, if not worse. Hence, the center of the cocked hat, though might give a reasonable value of log likelihood, won't guarantee maximum value for the log likelihood.

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:
Consider the loss matrix L where L_{kj} denotes the cost of classifying an input of true class C_k as C_j. Assuming number of classes is N, L is a matrix $\in \mathbb{R}^{N\times N}$.
The reject option gives the maximum loss threshold (ψ here), and if we obtain a loss $\geqslant \psi$ we don't classify the given input.
The expected loss for a such a classifier will be given as :

$$\mathbb{E}_{X,C_K}[L] = \int_x \cdots \int \sum_{k=1}^{N} p(x, C_k) \left(\sum_{j=1}^{N} L_{kj} \left(\mathbb{I}_{\{h(x)=j\}} \right) \right) dx$$

where $h(x)$ is the class number to which x is classified.
Since Bayes decision theory allows to classify each x independently, we write the above in a form to get the decision criteria :

$$\mathbb{E}_{X,C_K}[L] = \int_x \cdots \int \sum_{j=1}^{N} \left(\sum_{k=1}^{N} p(C_k \mid x) L_{kj} \right) \left(\mathbb{I}_{\{h(x)=j\}} \right) p(x)\, dx \tag{1}$$

To minimise the expected loss given by 1, we need to minimise the term inside the bigger brackets, which is $\mathbb{E}_{C_K}[L \mid X = x, h(X) = C_j]$ for each x independently, by choosing an appropriate j as $h(x)$.
Hence the decision criteria (including the reject option also) is given by :

$$h(x) = \begin{cases} \arg\min\limits_{C_j} \sum\limits_{k=1}^{N} p(C_k \mid x) L_{kj} & \text{if } \mathbb{E}_{X,C_K}[L] < \psi \\ \text{no decision} & \text{if } \mathbb{E}_{X,C_K}[L] \geqslant \psi \end{cases} \tag{2}$$

For the simpler case with a $0-1$ loss, where $L_{kj} = 1 - I_{kj}$ (with I_{kj} being 1 for $k = j$ and 0 otherwise), the expected loss given by equation 1 will becomes :

$$\begin{aligned} \mathbb{E}_{X,C_K}[L] &= \int_x \cdots \int \sum_{j=1}^{N} \left(\sum_{k=1}^{N} p(C_k \mid x) \left(\mathbb{I}_{\{k=j\}} \right) \right) \left(\mathbb{I}_{\{h(x)=j\}} \right) p(x)\, dx \\ &= \int_x \cdots \int \sum_{j=1}^{N} \left(\sum_{\substack{k=1 \\ k\neq j}}^{N} p(C_k \mid x) \right) \left(\mathbb{I}_{\{h(x)=j\}} \right) p(x)\, dx \end{aligned} \tag{3}$$

which leads to the simplified decision criteria for $0-1$ loss as follows :

$$h(x) = \begin{cases} \arg\min\limits_{C_j} \sum\limits_{\substack{k=1 \\ k\neq j}}^{N} p(C_k \mid x) L_{kj} & \text{if } \mathbb{E}_{X,C_k}[L] < \psi \\ \text{no decision} & \text{if } \mathbb{E}_{X,C_k}[L] \geqslant \psi \end{cases} \tag{4}$$

(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 $P(x|C_1 = 1) \sim N(-2, 1)$, $P(x|C_2 = 2) \sim N(0, 1)$ and $P(x|C_3) \sim N(2, 1)$. Find the Bayes classifier $h(x)$.

Solution:
We saw the decision criteria for a bayes classifier in part (a) of the question. Since we can decide for each x independently, multiplying the function which needs to be minimised by the predicted class by $p(x)$ is common for all won't change the criteria.
Hence we use the following decision criteria for this problem (assuming no reject option as nothing is mentioned) :

$$\begin{aligned} h(x) &= \arg\min_{C_j} \ \mathbb{E}_{C_k}\big[L \mid X = x,\, h(X) = C_j\big]\, p(x) \\ &= \arg\min_{j} \ \sum_{k=1}^{3} p(C_k \mid x)\, p(x) L_{kj} \\ &= \arg\min_{j} \ \sum_{k=1}^{3} p(C_k, x) L_{kj} \\ &= \arg\min_{j} \ < p(\cdot, x), L_{\cdot,j} > \end{aligned} \tag{1}$$

where $< u, v >$ represents the dot product between vectors u and v.
It is given that $p(C_k) = \frac{1}{3} \quad \forall\, k \in \{1, 2, 3\}$ and $p(C_k, x)$ can be obtained as $p(x)\, p(C_k \mid x)$.

$$p(\cdot, x) = \begin{bmatrix} \frac{1}{3} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+2)^2}{2}\right) \\ \frac{1}{3} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x)^2}{2}\right) \\ \frac{1}{3} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-2)^2}{2}\right) \end{bmatrix}$$

The σ for all class conditionals' gaussian distributions is 1 and the are μ's are -2, 0 and 2 respectively.
Let us find $\mathbb{E}_{C_k}\left[L \mid X = x,\, h(X) = C_j\right] p(x)$ for an arbitrary x, for different j's.
For predicting class as C_1 :

$$L_{\cdot,1} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$\mathbb{E}_{C_k}\left[L \mid X = x,\, h(X) = C_1\right] = \frac{1}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x)^2}{2}\right) + \frac{2}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-2)^2}{2}\right) \quad (2)$$

For predicting class as C_2 :

$$L_{\cdot,2} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

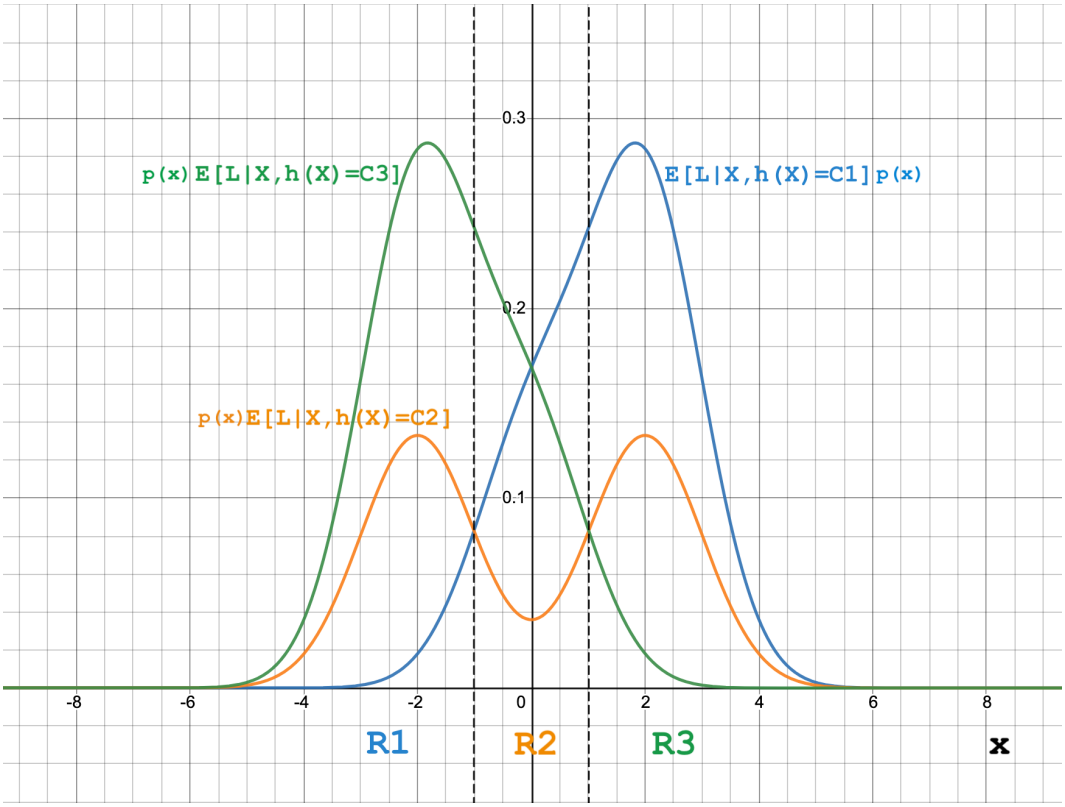
$$\mathbb{E}_{C_k}\left[L \mid X = x,\, h(X) = C_2\right] = \frac{1}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+2)^2}{2}\right) + \frac{1}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-2)^2}{2}\right) \quad (3)$$

For predicting class as C_3 :

$$L_{\cdot,3} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbb{E}_{C_k}\left[L \mid X = x,\, h(X) = C_3\right] = \frac{2}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+2)^2}{2}\right) + \frac{1}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x)^2}{2}\right) \quad (4)$$

The class prediction based on the value of j minimising $\mathbb{E}_{C_k}\left[L \mid X, h(X) = C_j\right] p(x)$ as obtained in equations 2, 3, 4 for different x, is shown in the figure below :



R1 - region where the Bayes classifier will predict the class as C_1
R2 - region where the Bayes classifier will predict the class as C_2
R3 - region where the Bayes classifier will predict the class as C_3
Calling $\mathbb{E}_{C_K}\big[L \mid X = x,\, h(X) = C_j\big]p(x)$ as $g_j(x)$ for simplicity, formally the Bayes classifier function is defined as :

$$h(x) = \begin{cases} C_1 & \text{if } \min\big(g_1(x), g_2(x), g_3(x)\big) = g_1(x) \\ C_2 & \text{if } \min\big(g_1(x), g_2(x), g_3(x)\big) = g_2(x) \\ C_3 & \text{if } \min\big(g_1(x), g_2(x), g_3(x)\big) = g_3(x) \end{cases}$$

where $g_1(x), g_2(x), g_3(x)$ are given by the equation 2, 3 and 4.

(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:

Assuming $0-1$ loss, the loss matrix for this case (with 2 classes) will look like :

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

Let us assume no reject option, as nothing is told about it.
For simplicity, let $g_j(x) = \mathbb{E}_{C_k}\big[L \mid X = x,\, h(X) = C_j\big]p(x)$.
As in part (b), $p(x)$ is common across all $g_j(x)$ for arbitrary x, hence multiplying $\mathbb{E}_{C_k}\big[L \mid X = x,\, h(X) = C_j\big]$ by $p(x)$ doesn't affect the decision criteria and for this case it will look like :

$$\begin{aligned}
h(x) &= \arg\min_{C_j} \; g_j(x) \\
&= \arg\min_{j} \; \sum_{k=1}^{2} p(C_k \mid x)\, p(x) L_{kj} \\
&= \arg\min_{j} \; \sum_{k=1}^{2} p(C_k, x) L_{kj} \\
&= \arg\min_{j} \; \sum_{k=1}^{2} p(C_k, x) \mathbb{I}_{\{k \neq j\}} \\
&= \arg\min_{j} \; p(C_{3-j}, x) \qquad \text{(assuming } C_1, C_2 \text{ are the classes)} \qquad (1)
\end{aligned}$$

It is given that prior probabilities of the classes are equal $\implies p(C_1) = p(C_2) = \frac{1}{2}$.
And, join distribution $p(C_k, x)$ is given by $p(x)p(x \mid C_k)$.

Continuing from equation 1,

$$
\begin{aligned}
h(x) &= \begin{cases} C_1 & \text{if } p(C_1, x) > p(C_2, x) \\ C_2 & \text{otherwise} \end{cases} \\
&= \begin{cases} C_1 & \text{if } \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right) > \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right) \\ C_2 & \text{otherwise} \end{cases} \\
&= \begin{cases} C_1 & \text{if } \exp\left(\frac{(x-\mu_2)^2 - (x-\mu_1)^2}{2\sigma^2}\right) > 1 \\ C_2 & \text{otherwise} \end{cases} \\
&= \begin{cases} C_1 & \text{if } \left(\frac{(x-\mu_2)^2 - (x-\mu_1)^2}{2\sigma^2}\right) > 0 \\ C_2 & \text{otherwise} \end{cases} \\
&= \begin{cases} C_1 & \text{if } 2x\mu_1 - 2x\mu_2 + \mu_2^2 - \mu_1^2 > 0 \\ C_2 & \text{otherwise} \end{cases}
\end{aligned}
\tag{4}
$$

From the above expression we can conclude that the decision boundary is :

$$
\begin{aligned}
& 2x\mu_1 - 2x\mu_2 + \mu_2^2 - \mu_1^2 = 0 \\
\implies & (\mu_1 - \mu_2)\big(2x - (\mu_2 + \mu_1)\big) = 0 \\
\implies & x = \frac{\mu_1 + \mu_2}{2}
\end{aligned}
$$

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 us introduce the latent variable z which is a 2-dimensional random variable having 1-of-N representation, taking two values $[1, 0]^T \equiv (z_1 = 1)$ and $[0, 1]^T \equiv (z_2 = 1)$ with probabilities λ and $(1 - \lambda)$ respectively.
Likelihood function for the two-component categorical mixture model to fit D_2 given by $(w_1, w_2, \cdots, w_k)$ is :

$$\begin{aligned}
\mathcal{L}(\lambda;\, \theta, \theta_1, D_2) &= \log P(D_2;\, \lambda, \theta, \theta_1) \\
&= \log \left(\prod_{i=1}^{k} p(w_i;\, \lambda, \theta, \theta_1) \right) \\
&= \log \left(\prod_{i=1}^{k} \Big(p(z_1 = 1;\, \lambda) p(w_i \mid z_1 = 1;\, \lambda, \theta, \theta_1) \right. \\
&\qquad \left. + p(z_2 = 1;\, \lambda) p(w_i \mid z_2 = 1;\, \lambda, \theta, \theta_1) \Big) \right) \\
&= \log \left(\prod_{i=1}^{k} \Big(p(z_1 = 1) p(w_i \mid \theta_1) + p(z_2 = 1) p(w_i \mid \theta) \Big) \right) \\
&= \sum_{i=1}^{k} \log \Big(\lambda\, p(w_i \mid \theta_1) + (1 - \lambda)\, p(w_i \mid \theta) \Big) \qquad (1)
\end{aligned}$$

We see that we obtain a summation within the log, which usually results in a non-convex function, and in the process of finding the maximum likelihood estimators by setting the derivatives to zero, we may end up with coupled equations, or not polynomial time solvable forms.
Hence we proceed to use an iterative algorithm called the **Expectation-Maximization** algorithm, to estimate the maximum likelihood solution efficiently and easily.
For simplicity, let $\mu(w)$ and $\mu_1(w)$ denote $p(w \mid \theta)$ and $p(w \mid \theta_1)$ respectively. (They are given)
Let us see what happens on differentiating $\mathcal{L}(\lambda;\, \theta, \theta_1, D_2)$ (for simplicity denoted as $\mathcal{L}(\lambda)$)

wrt λ for this case :

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left(\sum_{i=1}^{k} \log\left(\lambda\, \mu_1(w_i) + (1-\lambda)\, \mu(w_i)\right)\right) = 0$$

$$\implies \sum_{i=1}^{k} \frac{\mu_1(w_i) - \mu(w_i)}{\lambda\, \mu_1(w_i) + (1-\lambda)\, \mu(w_i)} = 0 \tag{2}$$

Let us group terms with same denominator (same word) together. Let $w^{(i)}$ be the i^{th} word in the vocabulary and k_i be number of times word $w^{(i)}$ occurs in D_2. $\sum_{i=1}^{M} \mu(w^{(i)}) = \sum_{i=1}^{M} \mu_1(w^{(i)}) = 1$. Let there be totally M distinct words in the vocabulary V (i.e) $M = |V|$. Equation 2 then,

$$\implies \sum_{i=1}^{M} \frac{k_i\left(\mu_1(w^{(i)}) - \mu(w^{(i)})\right)}{\lambda\, \mu_1(w^{(i)}) + (1-\lambda)\, \mu(w^{(i)})} = 0$$

$$\implies \sum_{i=1}^{M} k_i\left(\mu_1(w^{(i)}) - \mu(w^{(i)})\right)\left(\prod_{\substack{j=1 \\ j\neq i}}^{M}\left(\lambda\left(\mu_1(w^{(j)}) - \mu(w^{(j)})\right) + \mu(w^{(j)})\right)\right) = 0 \tag{3}$$

It is clear that equation 3 is a polynomial in λ.
Abel-Ruffini theorem states that there is no closed form solution to roots of polynomial of degree greater than or equal to 5. Hence, it is clear that it is hard to maximise the log likelihood directly and we need an **iterative approach** to solve the derivative set to zero equation.

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

Solution:
Let us use the EM algorithm for obtaining the maximum likelihood estimates for the problem.
Using Jensen's inequality we get a ELBO (Evidence lower bound) for $p(w_i; \lambda, \theta, \theta_1)$ (the

equation below is log likelihood for a single word) :

$$\begin{aligned}\mathcal{L}(\lambda \mid w_i) = \log p(w_i; \lambda, \theta, \theta_1) &= \log\left(\sum_z p(w_i, z; \lambda, \theta, \theta_1)\right) \\ &= \log\left(\sum_z Q_i(z) \cdot \frac{p(w_i, z; \lambda, \theta, \theta_1)}{Q_i(z)}\right) \\ &\geqslant \sum_z Q_i(z) \cdot \log\left(\frac{p(w_i, z; \lambda, \theta, \theta_1)}{Q_i(z)}\right)\end{aligned}$$

The total log likelihood is (as mentioned in part (a)) :

$$\mathcal{L}(\lambda; \theta, \theta_1, D_2) = \sum_{i=1}^{k} \log\Big(\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p(w_i \mid \theta)\Big) \tag{1}$$

E-step :
It maximises ELBO over $Q_i(z)$ by setting it to $p(z \mid w_i; \lambda^{(t)}, \theta, \theta_1) \quad \forall\, i \in \{1, 2, \cdots, k\}$ at the $(t+1)^{th}$ step ($\lambda^{(t)}$ is the value of parameter obtained at the end of t^{th} step or iteration).

$$\begin{aligned}p(z_1 = 1 \mid w_i; \lambda, \theta, \theta_1) &= \frac{p(z_1 = 1; \lambda) p(w_i \mid z_1 = 1; \lambda, \theta, \theta_1)}{p(z_1 = 1; \lambda) p(w_i \mid z_1 = 1; \lambda, \theta, \theta_1) + p(z_2 = 1; \lambda) p(w_i \mid z_2 = 1; \lambda, \theta, \theta_1)} \\ &= \frac{\lambda\, p(w_i \mid \theta_1)}{\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p(w_i \mid \theta)}\end{aligned} \tag{2}$$

$$\begin{aligned}p(z_2 = 1 \mid w_i; \lambda, \theta, \theta_1) &= \frac{p(z_2 = 1; \lambda) p(w_i \mid z_2 = 1; \lambda, \theta, \theta_1)}{p(z_1 = 1; \lambda) p(w_i \mid z_1 = 1; \lambda, \theta, \theta_1) + p(z_2 = 1; \lambda) p(w_i \mid z_2 = 1; \lambda, \theta, \theta_1)} \\ &= \frac{(1-\lambda)\, p(w_i \mid \theta)}{\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p(w_i \mid \theta)}\end{aligned} \tag{3}$$

Both equations 2 and 3 aren't required at the same time, since $p(z_1 = 1 \mid w_i; \lambda, \theta, \theta_1) + p(z_2 = 1 \mid w_i; \lambda, \theta, \theta_1) = 1$, and one can be calculated if the other is known easily.
So using the results from equations 2 and 2, update rule for $Q_i^{(t+1)}(z)$ looks like :

$$\begin{aligned}Q_i^{(t+1)}(z_1) &= \frac{\lambda^{(t)}\, p(w_i \mid \theta_1)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)} \\ Q_i^{(t+1)}(z_2) &= \frac{(1-\lambda^{(t)})\, p(w_i \mid \theta)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)}\end{aligned}$$
$$\forall\, i \in \{1, 2, \cdots, k\}$$

M-step :
It maximizes ELBO over the unknown parameter λ.

$$\begin{aligned}
\lambda^{(t+1)} &= \arg\max_{\lambda} \sum_{i=1}^{k} \sum_{z} Q_i^{(t+1)}(z) \cdot \log \left(\frac{p(w_i, z; \lambda, \theta, \theta_1)}{Q_i^{(t+1)}(z)} \right) \\
&= \arg\max_{\lambda} \sum_{i=1}^{k} \sum_{z} Q_i^{(t+1)}(z) \cdot \log p(w_i, z; \lambda, \theta, \theta_1) \\
&= \arg\max_{\lambda} \sum_{i=1}^{k} \mathbb{E}_{z \sim p(z|x; \lambda^{(t)}, \theta, \theta_1)} \log p(w_i, z; \lambda, \theta, \theta_1)
\end{aligned}$$

The simplification in the second step above occurs because $-\log Q_i^{(t+1)}$ comes outside along with the summations as a constant, not depending on λ, hence can be ignored.

$$\begin{aligned}
\lambda^{(t+1)} &= \arg\max_{\lambda} \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \log \left(\lambda \, p(w_i \mid \theta_1) \right) + Q_i^{(t+1)}(z_2) \cdot \log \left((1-\lambda) \, p(w_i \mid \theta) \right) \right) \\
&= \arg\max_{\lambda} \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \log \left(\lambda \right) + Q_i^{(t+1)}(z_2) \cdot \log \left(1-\lambda \right) \right) \\
&= \arg\max_{\lambda} \; g(\lambda) \qquad \text{(say)}
\end{aligned}$$

The simplification in the second step above occurs because of the terms coming outside as constant, not depending on λ being ignored.

We differentiate $g(\lambda)$ wrt λ and set the derivative to zero to maximize $g(\lambda)$.

$$\frac{d\, g(\lambda)}{d\lambda} = \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \frac{d\, \log(\lambda)}{d\lambda} + Q_i^{(t+1)}(z_2) \cdot \frac{d\, \log(1-\lambda)}{d\lambda} \right) = 0$$

$$\implies \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \frac{1}{\lambda} - Q_i^{(t+1)}(z_2) \cdot \frac{1}{1-\lambda} \right) = 0$$

$$\implies \sum_{i=1}^{k} \left((1-\lambda)\, Q_i^{(t+1)}(z_1) - \lambda\, Q_i^{(t+1)}(z_2) \right) = 0 \qquad \lambda \neq 0,\, 1$$

$$\implies \sum_{i=1}^{k} \left(\frac{(1-\lambda)\, \lambda^{(t)}\, p(w_i \mid \theta_1) - \lambda\, (1-\lambda^{(t)})\, p(w_i \mid \theta)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)} \right) = 0 \qquad \lambda \neq 0,\, 1$$

$$\implies \sum_{i=1}^{k} \lambda \left(\frac{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)} \right) = \sum_{i=1}^{k} \frac{\lambda^{(t)}\, p(w_i \mid \theta_1)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)}$$

$$(\lambda \neq 0,\, 1)$$

$$\implies \lambda = \frac{1}{k} \sum_{i=1}^{k} \frac{\lambda^{(t)}\, p(w_i \mid \theta_1)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)} \qquad \lambda \neq 0,\, 1$$

Hence,

$$\lambda^{(t+1)} = \frac{1}{k} \sum_{i=1}^{k} \frac{\lambda^{(t)}\, p(w_i \mid \theta_1)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)}$$

(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:
If θ is unknown, it becomes another parameter that needs to be estimated by the maximum likelihood approach through the EM-algorithm.
$p(w \mid \theta)$ is a function of the unknown parameter θ, call it as $p_\theta(w)$.
The log likelihood then becomes :

$$\mathcal{L}(\lambda, \theta;\, \theta_1, D_2) = \sum_{i=1}^{k} \log \left(\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p_\theta(w_i) \right) \tag{1}$$

E-step :

It maximises ELBO over $Q_i(z)$ by setting it to $p(z \mid w_i; \lambda^{(t)}, \theta^{(t)}, \theta_1) \quad \forall\, i \in \{1, 2, \cdots, k\}$ at the $(t+1)^{th}$ step ($\lambda^{(t)}$ and $\theta^{(t)}$ are the values of parameters obtained at the end of t^{th} step or iteration).

$$\begin{aligned} p(z_1 = 1 \mid w_i; \lambda, \theta, \theta_1) &= \frac{p(z_1 = 1; \lambda) p(w_i \mid z_1 = 1; \lambda, \theta, \theta_1)}{p(z_1 = 1; \lambda) p(w_i \mid z_1 = 1; \lambda, \theta, \theta_1) + p(z_2 = 1; \lambda) p(w_i \mid z_2 = 1; \lambda, \theta, \theta_1)} \\ &= \frac{\lambda\, p(w_i \mid \theta_1)}{\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p(w_i \mid \theta)} \\ &= \frac{\lambda\, p(w_i \mid \theta_1)}{\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p_\theta(w_i)} \end{aligned} \tag{2}$$

$$\begin{aligned} p(z_2 = 1 \mid w_i; \lambda, \theta, \theta_1) &= \frac{p(z_2 = 1; \lambda) p(w_i \mid z_2 = 1; \lambda, \theta, \theta_1)}{p(z_1 = 1; \lambda) p(w_i \mid z_1 = 1; \lambda, \theta, \theta_1) + p(z_2 = 1; \lambda) p(w_i \mid z_2 = 1; \lambda, \theta, \theta_1)} \\ &= \frac{(1-\lambda)\, p(w_i \mid \theta)}{\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p(w_i \mid \theta)} \\ &= \frac{(1-\lambda)\, p_\theta(w_i)}{\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p_\theta(w_i)} \end{aligned} \tag{3}$$

Both equations 2 and 3 aren't required at the same time, since $p(z_1 = 1 \mid w_i; \lambda, \theta, \theta_1) + p(z_2 = 1 \mid w_i; \lambda, \theta, \theta_1) = 1$, and one can be calculated if the other is known easily.
So using the results from equations 2 and 2, update rule for $Q_i^{(t)}(z)$ looks like :

$$\begin{aligned} Q_i^{(t+1)}(z_1) &= \frac{\lambda^{(t)}\, p(w_i \mid \theta_1)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta^{(t)})} \\ &= \frac{\lambda^{(t)}\, p(w_i \mid \theta_1)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p_{\theta^{(t)}}(w_i)} \\ Q_i^{(t+1)}(z_2) &= \frac{(1-\lambda^{(t)})\, p(w_i \mid \theta^{(t)})}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta^{(t)})} \\ &= \frac{(1-\lambda^{(t)})\, p_{\theta^{(t)}}(w_i)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p_{\theta^{(t)}}(w_i)} \end{aligned}$$

$$\forall\, i \in \{1, 2, \cdots, k\}$$

M-step :
It maximizes ELBO over the unknown parameters λ and θ on which the log likelihood depends on.

$$\begin{aligned}
\lambda^{(t+1)} &= \arg\max_{\lambda,\theta} \sum_{i=1}^{k} \sum_{z} Q_i^{(t+1)}(z) \cdot \log\left(\frac{p(w_i, z;\, \lambda, \theta, \theta_1)}{Q_i^{(t+1)}(z)}\right) \\
&= \arg\max_{\lambda,\theta} \sum_{i=1}^{k} \sum_{z} Q_i^{(t+1)}(z) \cdot \log p(w_i, z;\, \lambda, \theta, \theta_1) \\
&= \arg\max_{\lambda,\theta} \sum_{i=1}^{k} \mathbb{E}_{z\sim p(z|x;\lambda^{(t)},\theta^{(t)},\theta_1)} \log p(w_i, z;\, \lambda, \theta, \theta_1)
\end{aligned}$$

The simplification in the second step above occurs because $-\log Q_i^{(t+1)}$ comes outside along with the summations as a constant, not depending on λ or θ, hence can be ignored.

$$\begin{aligned}
\lambda^{(t+1)} &= \arg\max_{\lambda,\theta} \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \log\left(\lambda\, p(w_i \mid \theta_1)\right) + Q_i^{(t+1)}(z_2) \cdot \log\left((1-\lambda)\, p(w_i \mid \theta)\right)\right) \\
&= \arg\max_{\lambda,\theta} \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \log\left(\lambda\right) + Q_i^{(t+1)}(z_2) \cdot \log\left((1-\lambda)\, p_\theta(w_i)\right)\right) \\
&= \arg\max_{\lambda,\theta}\ g(\lambda, \theta) \qquad\qquad \text{(say)}
\end{aligned}$$

The simplification in the second step above occurs because of the terms coming outside as constant, not depending on λ or θ being ignored.
We do partial differentiation of $g(\lambda, \theta)$ wrt λ and θ separately and set the derivatives to zero to maximize $g(\lambda, \theta)$.
Setting partial derivatives wrt λ to zero :

$$\begin{aligned}
&\frac{\partial\, g(\lambda)}{\partial \lambda} = \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \frac{\partial\, \log\left(\lambda\right)}{\partial \lambda} + Q_i^{(t+1)}(z_2) \cdot \frac{\partial\, \log\left((1-\lambda)\, p_\theta(w_i)\right)}{\partial \lambda}\right) = 0 \\
&\implies \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \frac{1}{\lambda} - Q_i^{(t+1)}(z_2) \cdot \frac{p_\theta(w_i)}{(1-\lambda)\, p_\theta(w_i)}\right) = 0 \\
&\implies \sum_{i=1}^{k} \left(Q_i^{(t+1)}(z_1) \cdot \frac{1}{\lambda} - Q_i^{(t+1)}(z_2) \cdot \frac{1}{1-\lambda}\right) = 0
\end{aligned}$$

We get to the same equation as in part (b), hence the same estimate for $\lambda^{(t+1)}$.

$$\lambda^{(t+1)} = \frac{1}{k} \sum_{i=1}^{k} \frac{\lambda^{(t)}\, p(w_i \mid \theta_1)}{\lambda^{(t)}\, p(w_i \mid \theta_1) + (1-\lambda^{(t)})\, p(w_i \mid \theta)}$$

Similarly setting partial derivatives wrt θ to zero, we can get estimates for $\theta^{(t+1)}$. θ can also be a vector in which case the gradients should be set to zero, to obtain $\theta^{(t+1)}$ vector.

(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:
In Bayesian Inference, λ is a random variable, which has a prior distribution. Since, $\lambda \in [0, 1]$, we can try choosing a **Beta distribution** for λ, as it can incorporate many different possible curves for the prior distribution (including uniform) by varying the hyperparameters a and b.
(If we choose a uniform distribution for λ it is as good as not knowing what prior to choose, and the posterior $\propto$ prior $\cdot$ likelihood will simply be the likelihood itself, as prior is constant. Hence, maximizing the posterior will become same as maximizing likelihood, so MAP estimates and ML estimates will be the same)
The beta distribution is given by :

$$Beta(\lambda \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\lambda^{(a-1)}(1-\lambda)^{(b-1)}$$

The posterior will be given by :

$$\begin{aligned} P(\lambda \mid \theta, \theta_1, D_2, a, b) &\propto P(D_2 \mid \lambda, \theta, \theta_1) \cdot P(\lambda \mid a, b) \\ &\propto \left(\prod_{i=1}^{k} \Big(\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p(w_i \mid \theta) \Big) \right) \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\lambda^{(a-1)}(1-\lambda)^{(b-1)} \end{aligned}$$

We can estimate λ by maximizing the log likelihood of the posterior (the normalizing factor for the posterior is common and can be ignore while maximizing), leading to MAP (Maximum-a-posteriori) estimates.

$$\begin{aligned} \lambda_{MAP} &= \arg\max_{\lambda} \ \log P(\lambda \mid \theta, \theta_1, D_2, a, b) \\ &\equiv \arg\max_{\lambda} \ \sum_{i=1}^{k} \left(\log \Big(\lambda\, p(w_i \mid \theta_1) + (1-\lambda)\, p(w_i \mid \theta) \Big) \right) + \log \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \right) + \\ &\qquad (a-1)\log\lambda + (b-1)\log(1-\lambda) \\ &= \arg\max_{\lambda} \ g(\lambda) \qquad \text{(say)} \end{aligned}$$

$g(\lambda)$ is almost like a super set of the function we had to maximize for ML estimates, as it contains the same summation within log as in the Maximum likelihood case, and even log

terms without summation inside coming from prior distributions. Hence differentiating $g(\lambda)$ wrt λ and setting derivative to zero will again lead to a non-closed form complex expression (as we saw in part(a)) for λ because of the summation within log term. Hence again the MAP estimate cannot be computed efficiently or easily. The EM algorithm cannot be directly used here, since the target function contains terms other than the summation within log, which makes it difficult to realise the latent variables. Hence, other numeric or iterative approaches need to be used to maximize $g(\lambda)$.

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:
$\theta = \{w_0, w_1\}$. For a given x_i let $\widehat{y}_i$ indicate the linear relation $w_0 + w_1 x_i$.
So, $y_i = \widehat{y}_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
For a given x and θ, $\widehat{y}_i$ is also fixed, hence $y_i \sim \mathcal{N}(\widehat{y}_i, \sigma^2)$. Hence,

$$P(y_i \mid x_i, \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \widehat{y}_i)^2}{2\sigma^2}\right)$$

Let the training dataset be represented as $\{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x} = \{x_i\}_{i=1}^n$ and $\mathbf{y} = \{y_i\}_{i=1}^n$. The likelihood of the n i.i.d samples is then given by :

$$P(\mathbf{y} \mid \mathbf{x}, \theta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \widehat{y}_i)^2}{2\sigma^2}\right)$$

The likelihood $P(\mathbf{y} \mid \mathbf{x}, \theta, \sigma)$ is equivalent to $P(\mathbb{D} \mid \theta, \sigma)$ for maximizing because $P(\mathbf{x})$ is common among all set of parameters and is unaffected by them.
And the log likelihood of the model is :

$$\begin{aligned}
\mathcal{L}(\theta; \mathbb{D}, \sigma) &= \log P(\mathbf{y} \mid \mathbf{x}, \theta, \sigma) \\
&= \log \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \widehat{y}_i)^2}{2\sigma^2} \right) \right) \\
&= \sum_{i=1}^{n} \left(-\log(\sigma\sqrt{2\pi}) - \frac{(y_i - \widehat{y}_i)^2}{2\sigma^2} \right) \\
&= -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2
\end{aligned}$$

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

Solution:
The ML estimates are obtained by setting the partial derivatives of the log likelihood wrt the unknown parameters to zero, so that the log likelihood is maximized.
Here the unknown parameters are w_0 and w_1 ($\theta = \{w_0, w_1\}$).
Setting partial derivative of log likelihood wrt w_0 to zero :

$$\begin{aligned}
&\frac{\partial}{\partial w_0} \mathcal{L}(\theta; \mathbb{D}, \sigma) = 0 \\
&\implies \frac{\partial}{\partial w_0} \left(-n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \right) = 0 \\
&\implies -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial w_0} \Big(y_i - (w_0 + w_{1,\mathrm{ML}} x_i) \Big)^2 = 0 \\
&\implies \sum_{i=1}^{n} 2(y_i - w_{0,\mathrm{ML}} - w_{1,\mathrm{ML}} x_i) \cdot (-1) = 0 \qquad \text{(by chain rule)} \\
&\implies \sum_{i=1}^{n} w_{0,\mathrm{ML}} = \sum_{i=1}^{n} (y_i - w_{1,\mathrm{ML}} x_i) \\
&\implies w_{0,\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - w_{1,\mathrm{ML}} x_i) \qquad (1)
\end{aligned}$$

Setting partial derivative of log likelihood wrt w_1 to zero :

$$\frac{\partial}{\partial w_1}\mathcal{L}(\theta; \mathbb{D}, \sigma) = 0$$

$$\implies \frac{\partial}{\partial w_1}\left(-n\log\sigma - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \right) = 0$$

$$\implies -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial w_1}\Big(y_i - (w_{0,ML} + w_1 x_i)\Big)^2 = 0$$

$$\implies \sum_{i=1}^{n} 2(y_i - w_{0,ML} - w_{1,ML}x_i)\cdot(-x_i) = 0 \qquad \text{(by chain rule)}$$

$$\implies \sum_{i=1}^{n} w_{1,ML}x_i^2 = \sum_{i=1}^{n} x_i\,(y_i - w_{0,ML})$$

$$\implies w_{1,ML} = \frac{\sum_{i=1}^{n} x_i\,(y_i - w_{0,ML})}{\sum_{i=1}^{n} x_i^2} \tag{2}$$

Substituting 1 in 2,

$$w_{1,ML} = \frac{\sum_{i=1}^{n} x_i\left(y_i - \frac{1}{n}\sum_{i=1}^{n}(y_i - w_{1,ML}x_i)\right)}{\sum_{i=1}^{n} x_i^2}$$

$$\implies w_{1,ML} = \frac{\sum_{i=1}^{n} x_i\,(y_i - \overline{y}) + \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2 w_{1,ML}}{\sum_{i=1}^{n} x_i^2} \qquad \left(\text{define } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

$$\implies \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2\right) w_{1,ML} = \sum_{i=1}^{n} x_i\,(y_i - \overline{y})$$

$$\implies w_{1,ML} = \frac{\sum_{i=1}^{n} x_i\,(y_i - \overline{y})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2} \tag{3}$$

Substituting w_1 in 1,

$$w_{0,ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_{1,ML}x_i)$$

$$\implies w_{0,ML} = \overline{y} - w_1\overline{x} \qquad \left(\text{define } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$\implies w_{0,ML} = \overline{y} - \overline{x}\frac{\sum_{i=1}^{n} x_i\,(y_i - \overline{y})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2} \tag{4}$$

Equation 3 and 4 represent the close form expressions for ML estimate of w_1 and w_0 denoted by $w_{1,ML}$ and $w_{0,ML}$ respectively.

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

Solution:
If σ is unknown it becomes another parameter whose value for maximising the log likelihood needs to be estimated. This is done by solving the equations obtained by setting the partial derivatives of the log likelihood wrt to unknown parameters $\theta = \{w_0, w_1\}$ and σ to zero.
The log likelihood function becomes :

$$\begin{aligned}
\mathcal{L}(\theta, \sigma; \mathbb{D}) &= \log P(\mathbb{D} \mid \theta, \sigma) \\
&= \log \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \widehat{y}_i)^2}{2\sigma^2} \right) \right) \\
&= -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2
\end{aligned}$$

The log likelihood function will remain the same, except now it is a function of σ as well. While setting the partial derivatives of log likelihood $\mathcal{L}(\theta, \sigma; \mathbb{D})$ wrt w_0 and w_1 to zero, we hold σ constant, hence the values of ML estimates for w_0 and w_1 don't change. We mention them again for completeness :

$$w_{0,ML} = \overline{y} - \overline{x} \frac{\sum_{i=1}^{n} x_i (y_i - \overline{y})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2} \tag{1}$$

$$w_{1,ML} = \frac{\sum_{i=1}^{n} x_i (y_i - \overline{y})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2} \tag{2}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
The ML estimates $w_{0,ML}$ and $w_{1,ML}$ don't depend on σ, hence 1 and 2 are the final closed form expressions for them.

Setting partial derivative of log likelihood wrt σ to zero :

$$\begin{aligned}
&\frac{\partial}{\partial \sigma}\mathcal{L}(\theta, \sigma; \mathbb{D}) = 0 \\
&\implies \frac{\partial}{\partial \sigma}\left(-n \log \sigma - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \right) = 0 \\
&\implies -\frac{n}{\sigma_{ML}} + \frac{1}{\sigma_{ML}^3}\sum_{i=1}^{n}\Big(y_i - (w_{0,ML} + w_{1,ML}x_i)\Big)^2 = 0 \\
&\implies \sigma_{ML}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_{0,ML} - w_{1,ML}x_i)^2 \qquad (\sigma_{ML} \neq 0) \\
&\implies \sigma_{ML} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - w_{0,ML} - w_{1,ML}x_i)^2} \qquad (3)
\end{aligned}$$

where $w_{0,ML}, w_{1,ML}$ values in RHS of equation 3 are substituted from equations 1 and 2 to obtain σ_{ML} (ML estimate of σ).

(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:
$w_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $w_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ are independent. Hence,

$$\begin{aligned}
P(\theta \mid \mu_0, \sigma_0, \mu_1, \sigma_1) &= P(w_0, w_1 \mid \mu_0, \sigma_0, \mu_1, \sigma_1) = P(w_0 \mid \mu_0, \sigma_0) \cdot P(w_1 \mid \mu_1, \sigma_1) \\
&= \frac{1}{2\pi\,\sigma_0\,\sigma_1}\exp\left(-\frac{(w_0 - \mu_0)^2}{2\sigma_0^2} - \frac{(w_1 - \mu_1)^2}{2\sigma_1^2} \right)
\end{aligned}$$

Using Bayes theorem, the posterior probability of the joint set of unknown parameters is :

$$\begin{aligned}
P(w_0, w_1 \mid \mathbf{x}, \mathbf{y}, \mu_0, \sigma_0, \mu_1, \sigma_1, \sigma) &\propto P(\mathbf{y} \mid \mathbf{x}, w_0, w_1, \sigma) \cdot P(w_0, w_1 \mid \mu_0, \sigma_0, \mu_1, \sigma_1) \\
&\propto P(\mathbf{y} \mid \mathbf{x}, w_0, w_1, \sigma) \cdot P(w_0 \mid \mu_0, \sigma_0) \cdot P(w_1 \mid \mu_1, \sigma_1)
\end{aligned}$$

The parameters are set to the value that maximize the log of posterior probability (shown above), and we call this the **MAP (Maximum-a-Posteriori)** technique. w_0 and w_1 are the

parameters to be estimated here by the MAP method.

$$
\begin{aligned}
\{w_{0,\mathrm{MAP}}, w_{1,\mathrm{MAP}}\} &= \arg\max_{w_0,w_1} \ \log \mathrm{P}(w_0, w_1 \mid \mathbf{x}, \mathbf{y}, \mu_0, \sigma_0, \mu_1, \sigma_1, \sigma) \\
&\equiv \arg\max_{w_0,w_1} \ \log \mathrm{P}(\mathbf{y} \mid \mathbf{x}, w_0, w_1, \sigma) \cdot \mathrm{P}(w_0 \mid \mu_0, \sigma_0) \cdot \mathrm{P}(w_1 \mid \mu_1, \sigma_1) \\
&= \arg\max_{w_0,w_1} \ \log \mathrm{P}(\mathbf{y} \mid \mathbf{x}, w_0, w_1, \sigma) + \log \mathrm{P}(w_0 \mid \mu_0, \sigma_0) + \log \mathrm{P}(w_1 \mid \mu_1, \sigma_1) \\
&= \arg\max_{w_0,w_1} \ \left(-n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \right) \\
&\qquad + \left(-\log(2\pi\, \sigma_0\, \sigma_1) - \frac{(w_0 - \mu_0)^2}{2\sigma_0^2} - \frac{(w_1 - \mu_1)^2}{2\sigma_1^2} \right) \\
&= \arg\max_{w_0,w_1} \ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(y_i - (w_0 + w_1 x_i)\big)^2 - \frac{(w_0 - \mu_0)^2}{2\sigma_0^2} - \frac{(w_1 - \mu_1)^2}{2\sigma_1^2}
\end{aligned}
$$

In the last step, we have removed constants that won't affect the solution.
We maximise the function above by setting the partial derivatives wrt w_0 and w_1 to zero.
Setting partial derivative of log likelihood wrt w_0 to zero :

$$
\begin{aligned}
&\frac{\partial}{\partial w_0} \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(y_i - (w_0 + w_1 x_i)\big)^2 - \frac{(w_0 - \mu_0)^2}{2\sigma_0^2} - \frac{(w_1 - \mu_1)^2}{2\sigma_1^2} \right) = 0 \\
&\implies -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial w_0} \Big(\big(y_i - (w_0 + w_1 x_i)\big)^2 \Big) - \frac{\partial}{\partial w_0} \frac{(w_0 - \mu_0)^2}{2\sigma_0^2} - 0 = 0 \\
&\implies \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = \frac{w_0 - \mu_0}{\sigma_0^2} \\
&\implies w_0 \left(\frac{1}{\sigma_0^2} + \sum_{i=1}^{n} \frac{1}{\sigma^2} \right) = \frac{\mu_0}{\sigma_0^2} - \frac{w_1}{\sigma^2} \sum_{i=1}^{n} x_i + \frac{1}{\sigma^2} \sum_{i=1}^{n} y_i \\
&\implies w_0 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) = \frac{\mu_0}{\sigma_0^2} - \frac{w_1\, n\overline{x}}{\sigma^2} + \frac{n\overline{y}}{\sigma^2} \qquad \left(\text{define } \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i\right) \\
&\implies w_{0,\mathrm{MAP}} = \frac{\frac{\mu_0}{\sigma_0^2} + n\left(\frac{\overline{y}}{\sigma^2} - \frac{w_{1,\mathrm{MAP}} \overline{x}}{\sigma^2} \right)}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \qquad (1)
\end{aligned}
$$

(Not showing $w_{0,\mathrm{MAP}}, w_{1,\mathrm{MAP}}$ at every step to avoid cluttering the equation)

Setting partial derivative of log likelihood wrt w_1 to zero :

$$\frac{\partial}{\partial w_1}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-(w_0+w_1x_i)\right)^2-\frac{(w_0-\mu_0)^2}{2\sigma_0^2}-\frac{(w_1-\mu_1)^2}{2\sigma_1^2}\right)=0$$

$$\implies -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial w_1}\left(\left(y_i-(w_0+w_1x_i)\right)^2\right)-0-\frac{\partial}{\partial w_1}\frac{(w_1-\mu_1)^2}{2\sigma_1^2}=0$$

$$\implies \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-w_0-w_1x_i)\cdot x_i=\frac{w_1-\mu_1}{\sigma_1^2}$$

$$\implies w_1\left(\frac{1}{\sigma_1^2}+\sum_{i=1}^{n}x_i^2\frac{1}{\sigma^2}\right)=\frac{\mu_1}{\sigma_1^2}-\frac{w_0}{\sigma^2}\sum_{i=1}^{n}x_i+\frac{1}{\sigma^2}\sum_{i=1}^{n}y_ix_i$$

$$\implies w_1\left(\frac{1}{\sigma_1^2}+\frac{n\overline{x^2}}{\sigma^2}\right)=\frac{\mu_1}{\sigma_1^2}+\frac{n\overline{xy}}{\sigma^2}-\frac{w_0\,n\overline{x}}{\sigma^2}\qquad\left(\text{define }\overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2\text{ and }\overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i\right)$$

$$\implies w_{1,MAP}=\frac{\frac{\mu_1}{\sigma_1^2}+n\left(\frac{\overline{xy}}{\sigma^2}-\overline{x}\frac{w_{0,MAP}}{\sigma^2}\right)}{\frac{1}{\sigma_1^2}+\frac{n\overline{x^2}}{\sigma^2}}\tag{2}$$

Say, $\frac{1}{\sigma_1^2}+\frac{n\overline{x^2}}{\sigma^2}=A$ and $\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}=B$.
Substituting equation 1 in 2,

$$w_1=\frac{\frac{\mu_1}{\sigma_1^2}+n\left(\frac{\overline{xy}}{\sigma^2}\right)}{A}-\frac{n\overline{x}w_0}{A\sigma^2}$$

$$w_1=\frac{\frac{\mu_1}{\sigma_1^2}+n\left(\frac{\overline{xy}}{\sigma^2}\right)}{A}-\frac{n\overline{x}}{A\sigma^2}\cdot\frac{\frac{\mu_0}{\sigma_0^2}+n\left(\frac{\overline{y}}{\sigma^2}-\frac{w_1\overline{x}}{\sigma^2}\right)}{B}$$

$$w_1=\frac{\mu_1}{A\sigma^2}+\frac{n\overline{xy}}{A\sigma^2}-\frac{n\mu_0\overline{x}}{AB\sigma^2\sigma_0^2}-\frac{\overline{x}n^2\overline{y}}{AB\sigma^4}+\frac{n^2w_1\overline{x}^2}{AB\sigma^4}$$

$$w_{1,MAP}=\frac{\frac{\mu_1}{A\sigma^2}+\frac{n\overline{xy}}{A\sigma^2}-\frac{n\mu_0\overline{x}}{AB\sigma^2\sigma_0^2}-\frac{\overline{x}n^2\overline{y}}{AB\sigma^4}}{1-\frac{n^2\overline{x}^2}{AB\sigma^4}}\tag{3}$$

We obtained a closed form solution for $w_{1,MAP}$. Substituting value of $w_{1,MAP}$ back into 1, we get closed form solution for $w_{1,MAP}$ also.
We see that $w_{1,MAP}, w_{0,MAP}$ are similar in form to $w_{1,ML}, w_{0,ML}$ except it contains extra terms involving new hyperparameters $\mu_0, \sigma_0, \mu_1, \sigma_1$. Let's see the coupled equations we obtained for ML estimates in terms on the definitions used here :

$$w_{0,ML}=\frac{1}{n}\sum_{i=1}^{n}(y_i-w_{1,ML}x_i)$$
$$=\overline{y}-w_{1,ML}\overline{x}\tag{4}$$

$$
\begin{aligned}
w_{1,ML} &= \frac{\sum_{i=1}^{n} x_i \left(y_i - w_{0,ML}\right)}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{\overline{xy} - w_{0,ML}\overline{x}}{\overline{x^2}}
\end{aligned} \tag{5}
$$

Now let us see the coupled equations we obtained for MAP estimates, and set $\lim_{n\to\infty}$:

$$
\begin{aligned}
\lim_{n\to\infty} w_{0,MAP} &= \frac{\frac{\mu_0}{\sigma_0^2} + n\left(\frac{\overline{y}}{\sigma^2} - \frac{w_{1,MAP}\overline{x}}{\sigma^2}\right)}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \\
&= \frac{\frac{\mu_0}{n\sigma_0^2} + \left(\frac{\overline{y}}{\sigma^2} - \frac{w_{1,MAP}\overline{x}}{\sigma^2}\right)}{\frac{1}{n\sigma_0^2} + \frac{1}{\sigma^2}} \\
&= \overline{y} - w_{1,MAP}\overline{x}
\end{aligned}
$$

$$
\begin{aligned}
\lim_{n\to\infty} w_{1,MAP} &= \frac{\frac{\mu_1}{\sigma_1^2} + n\left(\frac{\overline{xy}}{\sigma^2} - \overline{x}\frac{w_{0,MAP}}{\sigma^2}\right)}{\frac{1}{\sigma_1^2} + \frac{n\overline{x^2}}{\sigma^2}} \\
&= \frac{\frac{\mu_1}{n\sigma_1^2} + \left(\frac{\overline{xy}}{\sigma^2} - \overline{x}\frac{w_{0,MAP}}{\sigma^2}\right)}{\frac{1}{n\sigma_1^2} + \frac{\overline{x^2}}{\sigma^2}} \\
&= \frac{\overline{xy} - w_{0,MAP}\overline{x}}{\overline{x^2}}
\end{aligned}
$$

We see that the coupled equations for MAP estimates reduces to those of ML estimates given by 4 and 5, when $n \to \infty$. And we showed that substituting on in the other, we can obtain closed form solution for the ML estimates in part(b). Following the same procedure, we will obtain the same closed form expressions for the MAP estimates as the ML estimates when allowing $n \to \infty$.

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 $X|Y=1 \sim \mathcal{N}(\mu_+, \Sigma_+)$

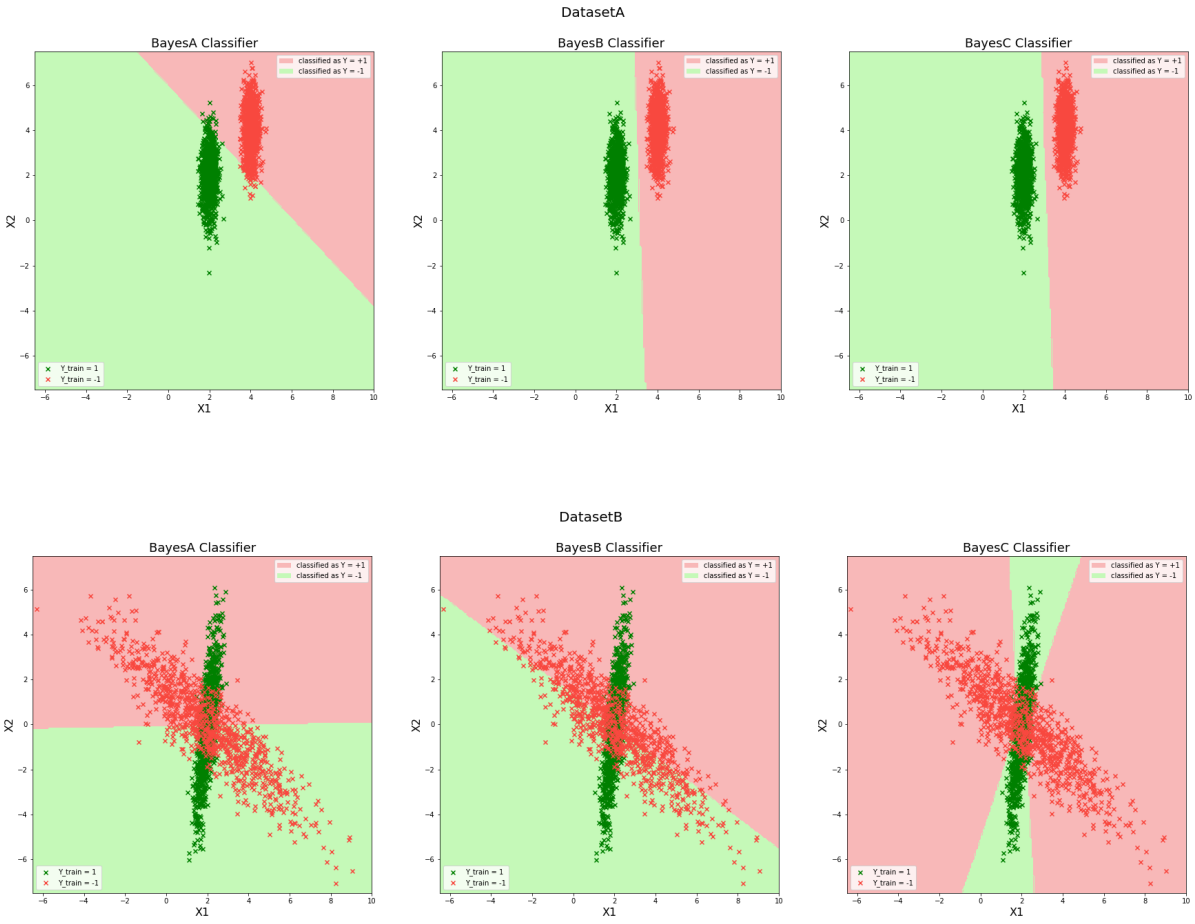
   Please see this folder for the template .ipynb file containing the helper functions, and you've to add

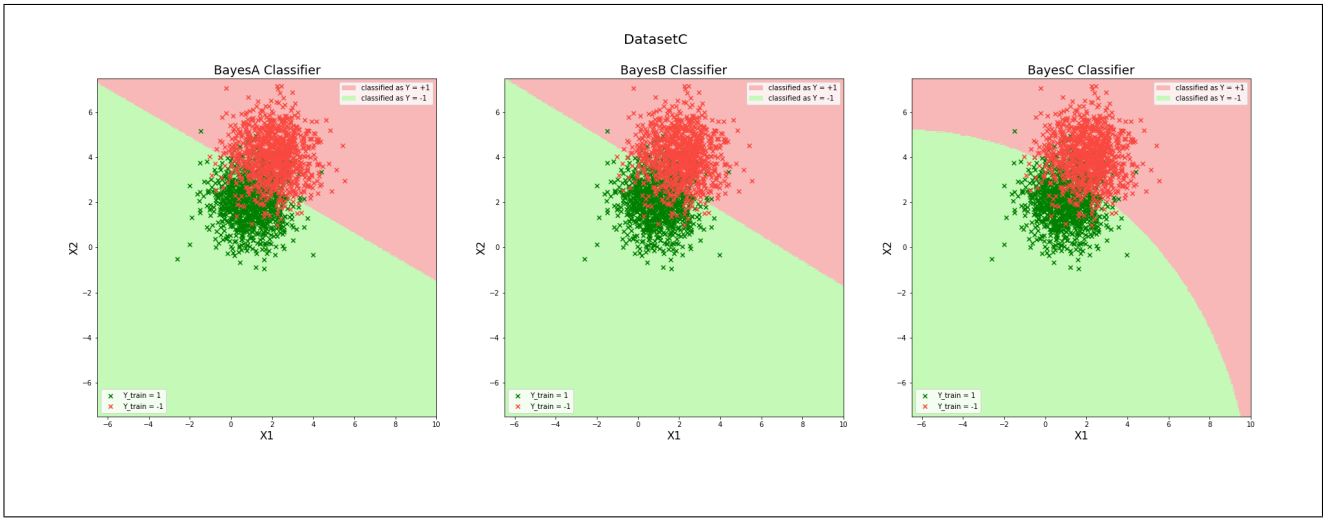
the missing code to this file (specifically, three functions function_for_A, function_for_B and function_for_C, and associated plotting/ROC code snippets) to implement the above three algorithms for the three datasets given in the same folder.

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).

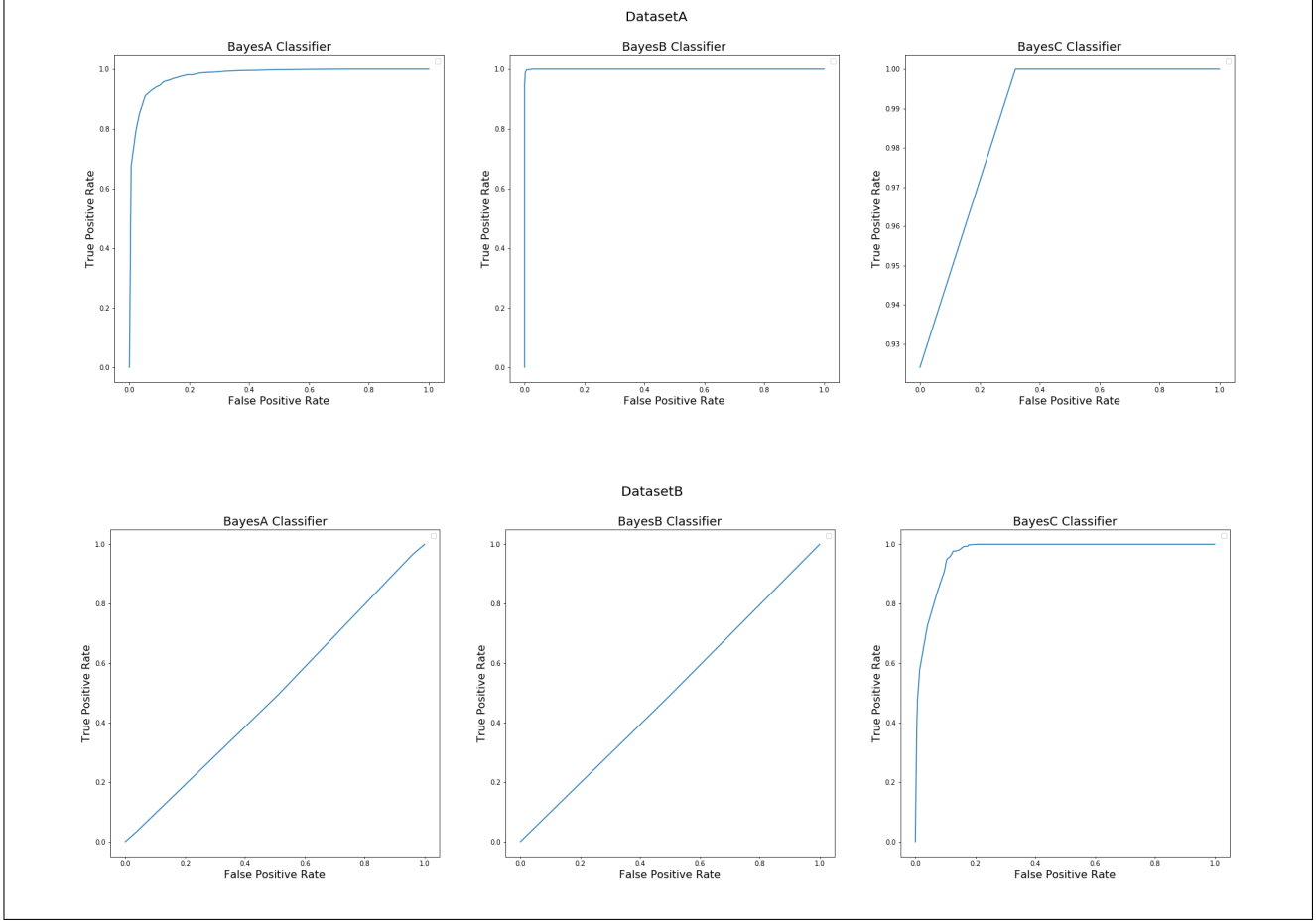
Solution:

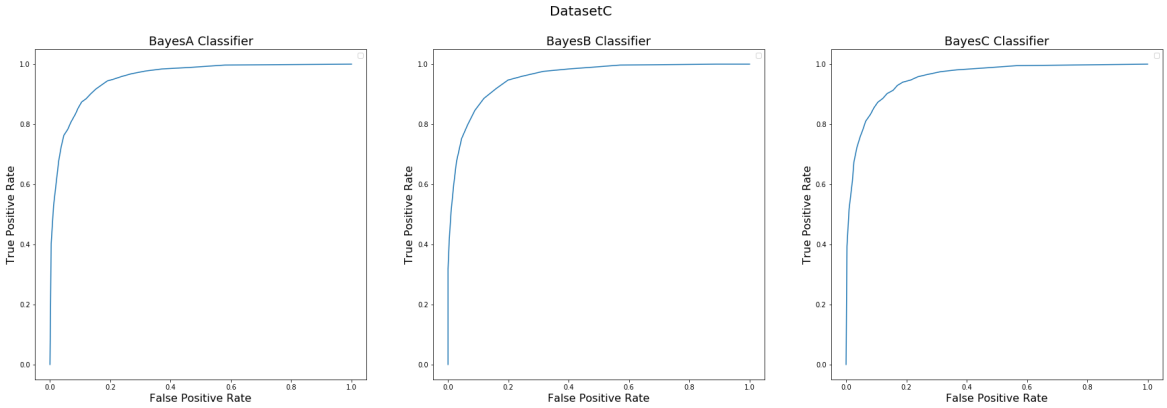




(b) (3 points) Give the ROC curves for all the classifiers. Note that a ROC curve plots the FPR (False Positive Rate) on the x-axis and TPR (True Positive Rate) on the y-axis. (9 plots)

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: The **Error Rates (in percentage)** calculated on the test datasets are shown in a table format below :

Dataset	BayesA	BayesB	BayesC
A	9.8	22.9	22.55
B	50.85	50.4	7.45
C	11.75	11.65	11.8

(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:
Bayes Classifier A :

- $X|Y = -1 \sim \mathcal{N}(\mu_-, I)$ and $X|Y = 1 \sim \mathcal{N}(\mu_+, I)$
- The decision boundary is linear in x ($w^T x + b$ form) and is given by :

$$2(\mu_- - \mu_+)^T x + (\mu_+^T \mu_+ - \mu_-^T \mu_-) + 2 \log \left(\frac{p(Y = -1)}{p(Y = +1)} \right) = 0$$

- The shape of decision boundary in the plots for classifier A is hence straight, representing the linear nature of the boundary.

Bayes Classifier B :

- $X|Y=-1 \sim \mathcal{N}(\mu_-, \Sigma)$ and $X|Y=1 \sim \mathcal{N}(\mu_+, \Sigma)$
- The decision boundary is again linear in x ($w^T x + b$ form) and is given by :

$$2(\mu_- - \mu_+)^T \Sigma^{-1} x + (\mu_+^T \Sigma^{-1} \mu_+ - \mu_-^T \Sigma^{-1} \mu_-) + 2\log\left(\frac{p(Y=-1)}{p(Y=+1)}\right) = 0$$

- The shape of decision boundary in the plots for classifier B is hence straight, representing the linear nature of the boundary.

Bayes Classifier C :

- $X|Y=-1 \sim \mathcal{N}(\mu_-, \Sigma_-)$ and $X|Y=1 \sim \mathcal{N}(\mu_+, \Sigma_+)$
- The decision boundary is quadratic in x (contains $x^T S x$ term along with linear terms) and is given by :

$$x^T(\Sigma_+^{-1} - \Sigma_-^{-1})x + 2(\mu_-^T \Sigma_-^{-1} - \mu_+^T \Sigma_+^{-1})x + (\mu_+^T \Sigma_+^{-1} \mu_+ - \mu_-^T \Sigma_-^{-1} \mu_-) + 2\log\left(\frac{p(Y=-1)}{p(Y=+1)}\right) + \log\left(\frac{|\Sigma_+^{-1}|}{|\Sigma_-^{-1}|}\right) = 0$$

- The shape of decision boundary in the plots for classifier C is hence curved, representing the second degree nature of the boundary.

Now, coming to the **General Observations :**

- The average error rates of the classifiers (from table c) are :
  - $avg\ error\ percentage(BayesA) = \frac{9.8+50.85+11.75}{3} = 24.133$
  - $avg\ error\ percentage(BayesB) = \frac{22.9+50.4+11.65}{3} = 28.3167$
  - $avg\ error\ percentage(BayesC) = \frac{22.55+7.45+11.8}{3} = 13.933$

  Hence the performance (in terms of correctness/accuracy in classifying) of the classifiers can be ranked approximately as : $B \lessapprox A << C$.
- The above result is due to the quadratic nature of the decision boundary of classifier C which allows more degrees of freedom and relatively complex decision boundaries.
- In terms of degrees of freedom in the parameters, $A < B < C$.

- Usually greater degrees of freedom might lead to overfitting where, the classifier fails to classify unseen test data accurately. We can see that happening to classifier B when compared with classifier A in dataset A (figure a), where it overfits the training data (decision boundary seems to separate the training data points so well), but fails in replicating the accuracy in test data, while classifier A gives a much lower error rate. In datasets B and C also, there is not much difference in error rates of classifiers A and B despite B being relatively more complex.
- Classifier C seems to gauge the situation better and gives rise to a good (almost) linear decision boundary for dataset A (as seen in a), despite being capable of producing curves boundaries.
- Classifiers A and B struggle a lot while classifying dataset B and give very high error rates, whereas classifier C is able to handle it well because of it's quadratic boundary. The ROC plots also indicate relatively how well classifier C is doing, where it is very near to ideal characteristics, while classifiers A and B give a poor ROC.
- There is a strong positive correlation between the error rates of the classifiers for a dataset and the corresponding ROC curves. Seeing one, we can qualitatively predict what to expect in the other.
- Non parametric approaches can be used to model complex distributions like (dataset B), but they are heavy on computation and storage, whereas parametric methods (used in this question) are easier to train and require only storing the parameters for classifying new data later. Parametric method also give good enough results when their degrees of freedom is increased, evident with the case of classifier C.