IITM-CS5691 : Pattern Recognition and Machine Learning
Assignment 1 (MKN Section)

Release Date: Sep 28, 2021
Due Date : Oct 13, 23:59

Roll No: ME18B181
Name: Aniruddha R Gandhewar
Roll No: ME18B182
Name: Vasudev Gupta
Team number: 1
References (if any):

1. (15 points) [DENSITY ESTIMATION]

   (a) (5 points) [PARAMETRIC MLE] Suppose that the lifetime of Philips brand light bulbs is modeled by an exponential distribution with (unknown) rate parameter λ or alternatively mean parameter μ. We test 6 bulbs and find they have lifetimes of 2, 6, 7, 1, 4, and 3 years, respectively. (i) (2 points) What is the MLE for λ and for μ, and (ii) (2 points) derive the bias of each of these estimators? (iii) (1 point) If the estimators are biased, how will you correct them to get unbiased estimators?

Solution:
(i) Lifetime of Philips brand light bulbs is modeled by an exponential distribution as follows:

$$p(x) = \lambda e^{-\lambda x}$$

or

$$p(x) = \frac{1}{\mu} e^{(-\frac{x}{\mu})}$$

where x is the lifetime of the bulb, λ is the rate parameter and, μ is the mean parameter. In order to estimate the parameters, we will use Maximum Likelihood Estimation (MLE).

Given: $D_N = \{2, 6, 7, 1, 4, 3\}$

For λ,

$$L(\lambda|D_N) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i} = \lambda^N e^{-\lambda(\sum_{i=1}^{N} x_i)}$$

Taking $\log_e$ on both sides,

$$\log L(\lambda|D_N) = N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

Since $\log x$ is a monotonically increasing function, the maxima of $f(x)$ and $\log f(x)$ occurs at the same x.

$$\frac{\partial \log L(\lambda|D_N)}{\partial \lambda} = 0$$

$$\frac{N}{\widehat{\lambda}_{ML}} = \sum_{i=1}^{N} x_i \;\Rightarrow\; \widehat{\lambda}_{ML} = \frac{N}{\sum_{i=1}^{N} x_i}$$

$$\boxed{\therefore\ \widehat{\lambda}_{ML} = 0.261}$$

Similarly for μ,

$$L(\mu|D_N) = \frac{1}{\mu^N} \exp(-\frac{\sum_{i=1}^{N} x_i}{\mu})$$

On solving for $\widehat{\mu}_{ML}$, we get,

$$\widehat{\mu}_{ML} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\boxed{\therefore\ \widehat{\mu}_{ML} = 3.833}$$

(ii) Suppose $X_1,\ X_2 \ldots X_N$ are i.i.d. random variables with probability density function $p(x) = \lambda e^{-\lambda x}$, then the maximum likelihood estimator for λ (also a random variable) is given as,

$$\widehat{\lambda}_{ML} = \frac{N}{\sum_{i=1}^{N} X_i}$$

We want to calculate $Bias\ [\widehat{\lambda}_{ML}]$,

$$Bias\ [\widehat{\lambda}_{ML}] = E\ [\widehat{\lambda}_{ML}] - \lambda$$

Now,

$$E\ [\widehat{\lambda}_{ML}] = E\ [\frac{N}{\sum_{i=1}^{N} X_i}] = N \times E\ [\frac{1}{Y}]$$

where $Y = \sum_{i=1}^{N} X_i$

Note: The Exponential distribution is a special case of the General Gamma distribution with two parameters, shape **a** and rate **b**. The probability density function of a Gamma Random Variable is:

$$p(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}},\ 0 < x < \infty$$

On taking $a = 1$ and $b = \frac{1}{\lambda}$, the Gamma distribution reduces to the Exponential distribution,

$$p(x) = \lambda e^{-\lambda x},\ 0 < x < \infty$$

Additive Property:

$$\text{If } X_i \sim Gamma(a_i, b)$$

then the random variable $Y = \sum_{i=1}^{N} X_i$ is also a Gamma random variable,

$$Y \sim Gamma(a^*, b^*)$$

$$where,\ a^* = \sum_{i=1}^{N} a_i,\ b^* = b$$

$$\therefore\ X_i \sim Gamma(1, \frac{1}{\lambda})\ \Rightarrow\ Y \sim Gamma(N, \frac{1}{\lambda})$$

Now,

$$E\left[Y^{-1}\right] = \int_0^\infty \frac{1}{y} \frac{\lambda^N}{\Gamma(N)} y^{N-1} e^{-\lambda y} dy = \frac{\lambda}{N-1} \int_0^\infty \frac{\lambda^{N-1}}{\Gamma(N-1)} y^{N-2} e^{-\lambda y} dy$$

$$E\left[Y^{-1}\right] = \frac{\lambda}{N-1}$$

Substituting $E\left[Y^{-1}\right]$ back into the equation for $E\left[\widehat{\lambda}_{ML}\right]$, we get,

$$E\left[\widehat{\lambda}_{ML}\right] = \frac{N}{N-1}\lambda$$

$$\boxed{\therefore Bias\left[\widehat{\lambda}_{ML}\right] = \frac{\lambda}{N-1} = \frac{\lambda}{5}}$$

$\widehat{\lambda}_{ML}$ is a biased estimator of λ and overestimates its value.

Similarly for $\widehat{\mu}_{ML}$,

$$\text{Bias}\,[\widehat{\mu}_{ML}] = \mathsf{E}\,[\widehat{\mu}_{ML}] - \mu$$

Now,

$$\mathsf{E}\,[\widehat{\mu}_{ML}] = \mathsf{E}\,[\frac{\sum_{i=1}^{N} X_i}{N}] = \frac{\sum_{i=1}^{N} \mathsf{E}\,[X_i]}{N}$$
$$\mathsf{E}\,[\widehat{\mu}_{ML}] = \mu$$

$$\boxed{\therefore \text{Bias}\,[\widehat{\mu}_{ML}] = 0}$$

$\widehat{\mu}_{ML}$ is an unbiased estimator of μ.

(iii) $\widehat{\lambda}_{ML}$ is a biased estimator of λ, while $\widehat{\mu}_{ML}$ is an unbiased estimator of μ. We can multiply $\widehat{\lambda}_{ML}$ by a factor of $\frac{N-1}{N}$ to get an unbiased estimate of λ.

$$\widehat{\lambda}_{ub} = \frac{N-1}{N}\,\widehat{\lambda}_{ML} = \frac{5}{6}\,\lambda$$

$$\therefore \text{Bias}\,[\widehat{\lambda}_{ub}] = \mathsf{E}\,[\widehat{\lambda}_{ub}] - \lambda = \frac{N-1}{N}\,\mathsf{E}\,[\widehat{\lambda}_{ML}] - \lambda = 0$$

(b) (5 points) [PARAMETRIC BAYESIAN] Assume we have following prior distribution on θ:

$$p(\theta) = \alpha\beta^{\alpha}\theta^{-\alpha-1}\mathbb{1}_{(\beta,\infty)}(\theta)$$

where $\mathbb{1}_{(\beta,\infty)}(\theta)$ is an indicator function which equals 1 when $\beta < \theta < \infty$ and 0 otherwise. $p(\theta)$ is called Pareto distribution which is denoted as $\theta \sim \text{Pareto}(\alpha, \beta)$.

i. ($1\frac{1}{2}$ points) Assume $\theta \sim \text{Pareto}(\alpha, \beta)$ and $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ which are conditionally independent given θ. What is the posterior distribution $p(\theta|D)$ where $D = (x_1, x_2, \ldots, x_n)$. Does it belong to any family of distributions that you recognize?

Solution:
Given that, $X_1, \ldots, X_n$ is observed data from Uniform$(0, \theta)$. We assume a prior for θ of the form,

$$\theta \sim \text{Pareto}(\alpha, \beta) = \alpha\beta^{\alpha}\theta^{-\alpha-1}\mathbb{1}_{(\beta,\infty)}(\theta)$$

where $\beta \geqslant x_{max}$, $x_{max} = max_{1 \leqslant i \leqslant n}\,\{x_i\}$. The likelihood of observing **n** such data points is given by,

$$\mathcal{L}_n(\theta) = \frac{1}{\theta^n}$$

Then, the posterior distribution $p(\theta|D_n)$, can be calculated using Bayes' theorem,

$$p(\theta|D_n) = \frac{p(D_n|\theta)\, p(\theta)}{p(D_n)}$$

- $p(D_n|\theta) = \mathcal{L}_n(\theta) = \frac{1}{\theta^n}$
- $p(\theta) = \alpha\beta^\alpha\theta^{-\alpha-1}\mathbb{1}_{(\beta,\infty)}(\theta)$
- $p(D_n) = \int_\theta p(D_n|\theta)\, p(\theta)d\theta = \int_\beta^\infty \alpha\beta^\alpha\theta^{-\alpha-n-1}d\theta = \frac{\alpha\beta^{-n}}{\alpha+n}$

On solving for the posterior distribution $p(\theta|D_n)$, we get,

$$\boxed{p(\theta|D_n) = (\alpha+n)\beta^{\alpha+n}\theta^{-\alpha-n-1}\mathbb{1}_{(\beta,\infty)}(\theta) \sim \text{Pareto}(\alpha+n,\ \beta)}$$

The posterior distribution $p(\theta|D_n)$ also belongs to the **Pareto** family of distributions.

ii. (1½ points) Using the above derived posterior, calculate the MAP estimate of θ? How does this compare to the MLE?

Solution:
Maximum A Posteriori (MAP),

$$\widehat{\theta}_{MAP} = \arg\max_\theta \text{Pareto}(\alpha+n,\ \beta)$$
$$\widehat{\theta}_{MAP} = \arg\max_\theta (\alpha+n)\beta^{\alpha+n}\theta^{-\alpha-n-1}\mathbb{1}_{(\beta,\infty)}(\theta)$$

Since, the Pareto distribution function is a monotonically decreasing function in θ,

$$\boxed{\therefore \widehat{\theta}_{MAP} = \beta}$$

Maximum Likelihood Estimate (MLE),

$$\widehat{\theta}_{MLE} = \arg\max_\theta \mathcal{L}_n(\theta)$$
$$\widehat{\theta}_{MLE} = \arg\max_\theta \frac{1}{\theta^n}$$

Since, the Likelihood function is a monotonically decreasing function in θ,

$$\boxed{\therefore \widehat{\theta}_{MLE} = x_{max} : max_{1\leqslant i\leqslant n}\{x_i\}}$$

Comparison between, the **MAP** and **MLE**,

$$\boxed{MAP(\beta) \geqslant MLE(x_{max} : max_{1\leqslant i\leqslant n}\{x_i\})}$$

iii. (2 points) Square loss is defined as $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. For the above derived posterior in (i), what estimator of θ minimizes the posterior expected square loss? Simplify your answer as much as possible. Is it the same as the MLE and/or the MAP?

Solution:
For the posterior expected square loss (Here, $\alpha^* = \alpha + n$),

$$E_\theta[L(\theta, \hat{\theta})] = \int_\beta^\infty (\theta - \hat{\theta})^2 \alpha^* \beta^{\alpha^*} \theta^{-\alpha^*-1} d\theta$$

$$E_\theta[L(\theta, \hat{\theta})] = \int_\beta^\infty \alpha^* \beta^{\alpha^*} \theta^{-\alpha^*+1} d\theta \; - \; 2\hat{\theta} \int_\beta^\infty \alpha^* \beta^{\alpha^*} \theta^{-\alpha^*} d\theta \; + \; \hat{\theta}^2 \int_\beta^\infty \alpha^* \beta^{\alpha^*} \theta^{-\alpha^*-1} d\theta$$

$$E_\theta[L(\theta, \hat{\theta})] = \frac{\alpha^* \beta^2}{\alpha^* - 2} \; - \; \frac{2\alpha^* \beta}{\alpha^* - 1} \hat{\theta} \; + \; \hat{\theta}^2$$

$$E_\theta[L(\theta, \hat{\theta})] = (\hat{\theta} - \frac{\alpha^* \beta}{\alpha^* - 1})^2 - (\frac{\alpha^* \beta}{\alpha^* - 1})^2 + \frac{\alpha^* \beta^2}{\alpha^* - 2}$$

Therefore, the estimator of θ that minimizes the posterior expected square loss is,

$$\boxed{\widehat{\theta}_L = \frac{\alpha^* \beta}{\alpha^* - 1}}$$

For large values of α^*, $\widehat{\theta}_L$ approaches the MAP, while for small values of α, it is significantly larger than the MAP.

(c) (5 points) [NON-PARAMETRIC METHOD] In class, we saw a Parzen window estimator using an unit hypercube as the Parzen window or kernel function; we will use an exponential kernel function here:

$$k(u) = \begin{cases} e^{-u} & u > 0, \\ 0 & u \leqslant 0. \end{cases}$$

If $D = \{x_1, x_2 \cdots x_n\}$ is a dataset of i.i.d. samples, each drawn from $U(0, 1)$, then,
(i) (3 points) show that the mean of the estimated density $p(x)$ is given by:

$$E_D[p(x)] = \begin{cases} 0 & x < 0 \\ 1 - e^{-\frac{x}{h}} & 0 \leqslant x \leqslant 1 \\ e^{\frac{1-x}{h}} - e^{-\frac{x}{h}} & x \geqslant 1. \end{cases}$$

Solution:
The estimated density $p(x)$,

$$p(x) = \frac{1}{nh} \sum_{1}^{n} k(\frac{x - x_i}{h})$$

where, x_i are i.i.d. samples, each drawn from $U(0, 1)$. We can now calculate the mean of the estimated density $E_D[p(x)]$.

Case 1: $\mathbf{x < 0}$

$$\frac{x - x_i}{h} < 0 \ \forall \ x < 0$$
$$\therefore k(\frac{x - x_i}{h}) = 0$$
$$\therefore p(x) = 0$$
$$\boxed{E_D[p(x)] = 0}$$

Case 2: $\mathbf{0 \leqslant x < 1}$

$$p(x) = \frac{1}{nh} \sum_{1}^{n} exp\left\{-(\frac{x - x_i}{h})\right\} \mathbb{1}_{x - x_i > 0}$$

Equivalently,

$$p(x) = \frac{1}{nh} \sum_{1}^{n} exp\left\{-(\frac{x - x_i}{h})\right\} \ where, \ x_i \sim U(0, x)$$

$$\therefore E_D[p(x)] = \frac{1}{nh} \sum_{1}^{n} exp\left\{-\frac{x}{h}\right\} E[exp\left\{\frac{X_i}{h}\right\}]$$

For $E[exp\left\{\frac{X_i}{h}\right\}]$, let $Y = exp\left\{\frac{X_i}{h}\right\}$,

$$P(Y \geqslant y) = P(exp\left\{\frac{X_i}{h}\right\} \geqslant y)$$
$$= P(X_i \geqslant h \ln y)$$
$$F_Y(y) = h \ln y$$
$$But, \ f_Y(y) = \frac{dF}{dy}$$
$$\therefore \ f_Y(y) = \frac{h}{y}$$

Now,

$$E[Y] = \int_{1}^{exp\{\frac{x}{h}\}} y \ \frac{h}{y} \ dy = h \ (exp\left\{\frac{x}{h}\right\} - 1)$$

$$\therefore\ E_D[p(x)] = \frac{1}{nh}\sum_{1}^{n} exp\,\{-\frac{x}{h}\}\,h\,(exp\,\{\frac{x}{h}\}-1)]$$

On simplifying, we get,

$$\boxed{\therefore\ E_D[p(x)] = 1 - e^{-\frac{x}{h}}}$$

Case 3: $\mathbf{x} \geqslant \mathbf{1}$: Since $\frac{x-x_i}{h} > 0\ \forall\ x_i$,

$$\therefore p(x) = \frac{1}{nh}\sum_{1}^{n} exp\,\{-(\frac{x-x_i}{h})\}$$

$$\therefore E_D[p(x)] = \frac{1}{nh}\sum_{1}^{n} exp\,\{-\frac{x}{h}\}\,E[exp\,\{\frac{X_i}{h}\}]$$

Here,

$$E[exp\,\{\frac{X_i}{h}\} = h\,(exp\,\{\frac{1}{h}\}-1)$$

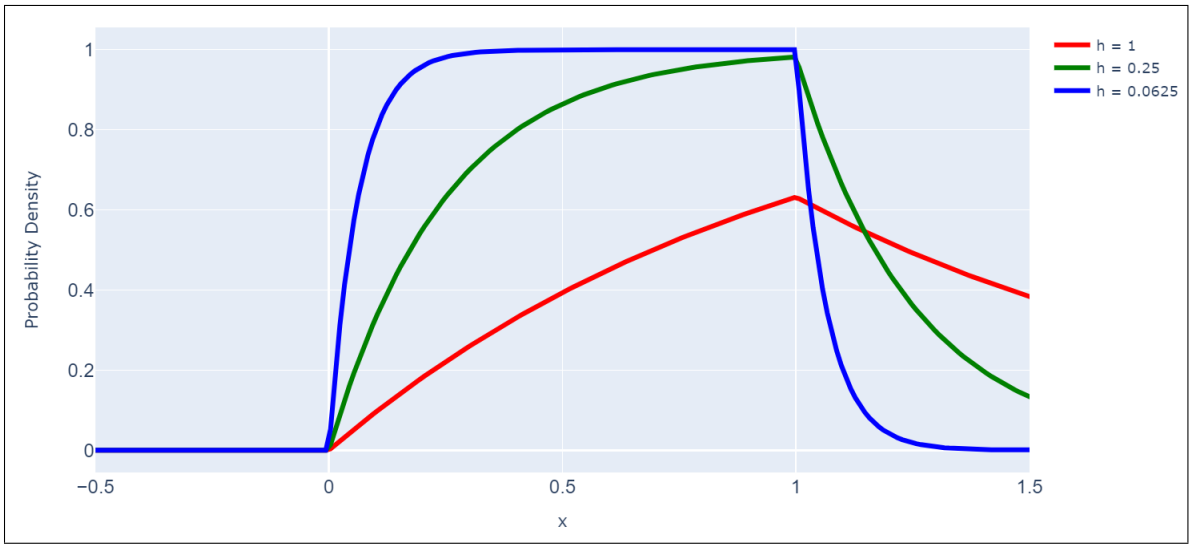
On simplifying, we get,

$$\boxed{\therefore\ E_D[p(x)] = e^{\frac{1-x}{h}} - e^{-\frac{x}{h}}}$$

Q.E.D.

(ii) (2 points) Also, plot $E_D[p(x)]$ vs x for different values of h ($h = 1, 0.25,$ and 0.0625). What do you observe?

Solution:



The lower the value of the bandwidth or smoothing parameter **h**, the better is the approximation of $U(0,\ 1)$.

2. (10 points) [BAYESIAN DECISION THEORY]

(a) (5 points) [Optimal Classifier by Pen/Paper] Let L be the loss matrix defined by $L = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$, where L_{ij} indicates the loss for an input x with i being the true class and j the predicted class. Given the data:

x	-2.9	1.4	0.4	-0.3	-0.7	0.9	1.8	0.8	-2.4	-1.4	1.2	2.3	2.8	-3.4
y	1	3	2	2	1	3	3	2	1	1	2	3	3	1

find the optimal Bayes classifier $h(x)$, and provide its decision boundaries/regions.

Solution:
Posterior distribution for the Bayes' Classifier is given as,

$$P(C_k|x) = \frac{p(x|C_k)P(C_k)}{\sum_j p(x|C_j)P(C_j)}$$

In the case of multi-class classification with Gaussian class-conditional probability density functions and multinoulli priors, the posterior distribution simplifies to,

$$P(C_k|x) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

where (for the case of uni-variate x),

$$a_k(x) = w_k x + w_{k0}$$
$$w_k = \frac{\mu_k}{\sigma^2}$$
$$w_{k0} = -\frac{\mu_k^2}{2\sigma^2} + \ln P(C_k)$$

Here, the parameters μ_k and σ^2 are calculated using Maximum Likelihood Estimation,

$$\mu_k = \frac{1}{N_k} \sum_{x_i \in C_k} x_i$$
$$\sigma^2 = \sum_i \frac{N_i}{N} s_i, \ \ where\ s_i = \frac{1}{N_i} \sum_{x_j \in C_i} (x_j - \mu_i)^2$$

Let's now calculate the posterior distribution for each of the three classes,

$$N = 14,\ N_1 = 5,\ N_2 = 4,\ N_3 = 5$$
$$\mu = 0.5,\ \mu_1 = -2.16,\ \mu_2 = 0.525,\ \mu_3 = 1.84$$
$$s_1 = 0.9704,\ s_2 = 0.3069,\ s_3 = 0.4424$$
$$\sigma^2 = \sum_i \frac{N_i}{N} s_i = 0.5923$$

Also,

$$P(C_1) = 0.357,\ P(C_2) = 0.286,\ P(C_3) = 0.357$$

For Class 1,

$$w_{10} = -4.968$$
$$w_1 = -3.647$$
$$a_1 = w_1 x + w_{10} = -3.647x - 4.968$$

For Class 2,

$$w_{20} = -1.484$$
$$w_2 = 0.886$$
$$a_2 = w_2 x + w_{20} = 0.886x - 1.484$$

For Class 3,

$$w_{30} = -3.888$$
$$w_3 = 3.107$$
$$a_3 = w_3 x + w_{30} = 3.107x - 3.888$$

Therefore, the posterior distribution functions,

$$P(1|x) = \frac{exp(-3.65x - 4.97)}{\sum_j exp(a_j)}, \quad P(2|x) = \frac{exp(0.89x - 1.48)}{\sum_j exp(a_j)}, \quad P(3|x) = \frac{exp(3.11x - 3.89)}{\sum_j exp(a_j)}$$

Now, given the Loss Matrix L, we want to minimize the expected loss,

$$E[L] = \int_x \sum_t L_{t,h(x)}\, p(x,t)\, dx$$

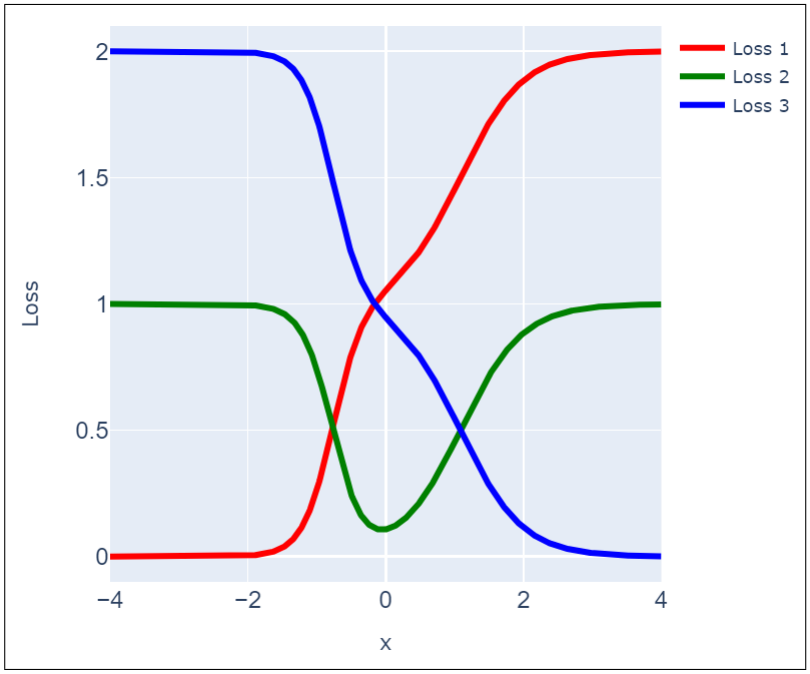
$$E[L] = \int_x \sum_t (P(t|x)\, L_{t,h(x)})\, p(x)\, dx$$

$$\therefore h(x) = \arg\min_j L^T [P(t|x)]_{j=1}^3$$

$$h(x) = \arg\min_j \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} P(1|x) \\ P(2|x) \\ P(3|x) \end{bmatrix}$$

$$h(x) = \arg\min_j \begin{bmatrix} P(2|x) + 2P(3|x) \\ P(1|x) + P(3|x) \\ 2P(1|x) + P(2|x) \end{bmatrix}$$

Plotting the loss, we get,



Decision Boundary of Bayes classifier $h(x)$,

$$\begin{aligned} &\text{Class } \mathbf{1}, \text{ if } -\infty < x < -0.77 \\ &\text{Class } \mathbf{2}, \text{ if } -0.77 < x < +1.08 \\ &\text{Class } \mathbf{3}, \text{ if } +1.08 < x < +\infty \end{aligned}$$

(b) (5 points) Consider the problem of classifying a pattern x into one of the k classes $c = 1, 2, ...k$. Assume that we have two different tests to determine the class to be assigned to pattern x. Test 1 assigns x to the class that maximizes the posterior probability, whereas test 2 to a class chosen based on randomized decision rule.

Test 1: $H_1(x) = c^* = \text{argmax}_c\, p(c|x)$
Test 2: $H_2(x) = c \sim p(c|x)$, where c is chosen based on the distribution $P(c = i|x)$ in a random fashion.

i. (1 point) Calculate the risk R_1 associated with test 1 in terms of the posterior probability using the zero-one loss function.

Solution:
The zero-one loss function is given by:

$$L(D_i|C_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

The risk is given by:

$$\begin{aligned} R_1(D_i|x) &= \sum_{i \neq j} P(C_j|x) \\ &= 1 - P(C_i|x) \end{aligned}$$

where D_i is the decision and C_i is the true class label.
The overall risk R_1 is,

$$R_1 = \int_x R_1(D_i|x)p(x)dx$$

According to Bayes Decision Theory,

$$D_i = \arg\max_{C_n} P(C_n|x)$$

$$\therefore\ R_1 = \int_x (1 - \max_j\{P(C_j|x)\})p(x)dx$$

ii. (2 points) Calculate the risk R_2 associated with test 2 in terms of the posterior probability using the zero-one loss function.

Solution:
The risk $R_2(D_i|x)$ is defined as,

$$R_2(D_i|x) = P_i(C_i|x)(1 - P_i(C_i|x))$$

The overall risk is,

$$\begin{aligned} R_2 &= \int_x (\sum_{i=1}^{k} P_i(C_i|x)(1 - P_i(C_i|x)))p(x)dx \\ &= \int_x (\sum_{i=1}^{k} P_i(C_i|x) - P_i^2(C_i|x)))p(x)dx \end{aligned}$$

On simplifying, we get,

$$\boxed{R_2 = \int_x (1 - \sum_{i=1}^{k} P_i^2(C_i|x)))p(x)dx}$$

iii. (2 points) Which test do you think would perform better always based on the risks R_1 and R_2? Also, specify the conditions under which both the tests behave the same.

Solution:
Based on the risks R_1 and R_2, Test 1 would perform better than Test 2.

$$\begin{aligned} \sum P^2(C_i|x) &\leqslant \sum P(C_i|x) \max\{P(C_i|x)\} \\ \sum P^2(C_i|x) &\leqslant \max\{P(C_i|x)\} \sum P(C_i|x) \\ \sum P^2(C_i|x) &\leqslant \max\{P(C_i|x)\} \end{aligned}$$

Multiplying both sides by -1,

$$-\sum P^2(C_i|x) \geqslant -\max\{P(C_i|x)\}$$

Adding 1 on both sides,

$$1 - \sum P^2(C_i|x) \geqslant 1 - \max\{P(C_i|x)\}$$

$$\Rightarrow \int_x (1 - \sum_{i=1}^{k} P_i^2(C_i|x)))p(x)dx \geqslant \int_x (1 - \max_j\{P(C_j|x)\})p(x)dx$$

$$\boxed{\therefore\ R_2 \geqslant R_1}$$

Hence, Test 1 would perform better than Test 2.
Both the tests would perform the same if the posterior distribution $P(C_i|x) = p\ \forall\ i\ \forall\ x$.
In such a scenario, randomly assigning classes is as good a strategy.

3. (15 points) [Linear regression]

   (a) (5 points) Say we have a linear regression dataset where every training datapoint $\{x_n, y_n\}$ has a weight q_n ($q_n > 0$) identified with it. Then we have the weighted error function (sum of squares) given by:

$$E_q(w) = \sum_{n=1}^{N} \frac{q_n(t_n - w^T x_n)^2}{2}.$$

   Derive the closed form solution for the minimizer w^* of this function. Express it in matrix format for a simplified expression.

Solution:
The weighted error function is given by:

$$E_q(w) = \sum_{n=1}^{N} \frac{q_n(t_n - w^T x_n)^2}{2}$$

Re-writing the above equation in matrix format,

$$E_q(w) = \frac{1}{2}\|Q(t - Xw)\|^2$$

where, $Q = diag(q_1,\ q_2,\ \ldots q_n)$, $X = [x_1,\ x_2,\ \ldots x_n]^T$ and $t = [t_1,\ t_2,\ \ldots t_n]^T$

$$\nabla_w E_q(w) = X^T Q(Xw - t)$$
$$0 = X^T Qt - X^T QXw^*$$

On Solving for w^*, we get,

$$\boxed{\therefore\ w^* = (X^T QX)^{-1} X^T Qt}$$

   (b) (5 points) We saw in class that the error function in case of ridge regression is given by:

$$\widetilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2 + \frac{\lambda}{2}w^T w.$$

Show that this error function is convex and is minimized by:

$$w^* = (\lambda I + \phi^T \phi)^{-1} \phi^T t.$$

Also show that $(\lambda I + \phi^T \phi)$ is invertible for any $\lambda > 0$.

Solution:
Error function $\widetilde{E}(w)$ is given by:,

$$\widetilde{E}(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2 + \frac{\lambda}{2} w^T w.$$

Re-writing in matrix format,

$$\widetilde{E}(w) = \frac{1}{2} \|t - \Phi w\|^2 + \frac{\lambda}{2} w^T w$$

In order to prove its convexity, we will show that its Hessian $(\nabla_w^2 \widetilde{E}(w))$ is a Positive Definite Matrix.

$$\nabla_w \widetilde{E}(w) = \Phi^T (\Phi w - t) + \lambda w$$
$$\nabla_w \widetilde{E}(w) = \Phi^T \Phi w - \Phi^T t + \lambda w$$

The Hessian is given as,

$$\nabla_w^2 \widetilde{E}(w) = \Phi^T \Phi + \lambda I$$

Now,

$$\text{Eigenvalues } (\Phi^T \Phi) = \mu_i, \ \mu_i \geqslant 0 \ \forall \ i$$
$$\text{Eigenvalues } (\Phi^T \Phi + \lambda I) = \mu_i + \lambda, \ \mu_i \geqslant 0 \ \forall \ i, \ \lambda > 0$$

Therefore,

$$\mu_i + \lambda > 0 \ \forall \ i \ \Rightarrow (\Phi^T \Phi + \lambda I) = \nabla_w^2 \widetilde{E}(w) \text{ is Positive Definite}$$

$$\therefore \ \widetilde{E}(w) \text{ is a } \textbf{convex} \text{ function.}$$

For w^*,

$$\nabla_w \widetilde{E}(w) = 0$$
$$\Phi^T (\Phi w^* - t) + \lambda w^* = 0$$

$$\boxed{\therefore \ w^* = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T t}$$

T.P.T: $(\lambda I + \phi^T \phi)$ is invertible for any $\lambda > 0$.
We earlier showed that,

$$\text{Hessian } \widetilde{E}(w) = \nabla_w^2 \widetilde{E}(w) = (\Phi^T \Phi + \lambda I) \text{ is Positive Definite for } \lambda > 0$$
$$\text{Positive Definiteness } \Rightarrow \text{ All Eigenvalues} > 0$$
$$\because (\Phi^T \Phi + \lambda I) \text{ has } \textbf{non} - \textbf{zero} \text{ eigenvalues } \Rightarrow (\Phi^T \Phi + \lambda I) \text{ is } \textbf{Invertible}.$$

Q.E.D.

(c) (5 points) Given a dataset

$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} \quad t = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

find all minimizers w of $E(w) = \frac{1}{2}\|Xw - t\|^2$, and indicate the one with the smallest norm. How does your answer change if you are looking for minimizers of $\widetilde{E}(w)$ instead (assuming $\lambda = 1$)?

Solution:
Let's first find the set of all w's which minimize the objective function $E(w)$.

$$\nabla_w E(w) = 0$$
$$X^T(Xw - t) = 0$$
$$X^T X w = X^T t$$

Substituting X and t, we get,

$$\begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$
$$\begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$
$$\therefore \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 3a - 1 \\ a \end{bmatrix}$$

Let's now find w with the smallest norm,

$$\|w\|^2 = (3a - 1)^2 + a^2$$
$$\frac{d\|w\|^2}{da} = 20a - 6$$
$$0 = 20a - 6$$
$$a = 0.3$$

w with the smallest norm is,

$$\boxed{w^* = \begin{bmatrix} -0.1 \\ 0.3 \end{bmatrix}}$$

Now, let's calculate w, if the objective function is $\widetilde{E}(w)$ instead (with $\lambda = 1$),

$$w^* = (X^T X + I)^{-1} X^T t$$

Solving for w_R^*, we get,

$$\boxed{w_R^* = \begin{bmatrix} -0.098 \\ 0.294 \end{bmatrix}}$$

4. (5 points) [Kernel methods] Let K_1,K_2 be two arbitary valid kernel functions mapping vectors from $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. For each of the cases below, show if it is a valid kernel or not with supporting arguments. (Hint: Keep your solutions brief by using earlier parts of this question to solve later parts whenever possible.)

   (a) (1 point) $K_3(x, y) = K_1(x, y) + K_2(x, y) + 7.5$

Solution: K_3 is a valid kernel.
Proof:
Defining K' as

$$K'(x, y) = K_1(x, y) + K_2(x, y)$$

Checking if K_3 is symmetrical:

$$K'(x, y) = K_1(y, x) + K_2(y, x)$$

where $K_1(y, x) = K_1(x, y)$ and $K_2(y, x) = K_2(x, y)$ as K_1 and K_2 are both valid kernels. Also, we can say that

$$K'(y, x) = K_1(y, x) + K_2(y, x)$$

Hence,

$$K'(y, x) = K'(x, y)$$

Hence, K′ is a symmetrical. Checking whether K′ is positive semi-definite

$$\begin{aligned} & u^T K' u \\ &= u^T (K_1 + K_2) u \\ &= u^T K_1 u + u^T K_2 u \\ &= \quad \geqslant 0 \qquad \geqslant 0 \\ &\geqslant 0 \end{aligned}$$

Hence, K′ is positive semi-definite also. We can conclude that K′ is a valid kernel. Now,

$$K_3(x, y) = K'(x, y) + 7.5$$

Let ϕ' denote the feature map of K′. The using the feature map $\phi' : x -> [\phi'(x), \sqrt{7.5}]$, we have

$$\phi(x)^T \phi(x) = \phi'(x)^T \phi'(x) + 7.5 = K_3(x, y)$$

Hence, $K_3(x, y)$ is a valid kernel

(b) (1 point) $K_4(x, y) = K_1(x, y)K_2(x, y)$ (product of two kernels)

Solution: K_4 is a valid kernel.
Proof:
Expressing $K_1(x, y)$ and $K_2(x, y)$ in terms of transformation function **a** and **b** respectively, we get the following:

$$\begin{aligned} K_1(x, y) &= a(x)^T a(y); & a(z) &= [a_1(z), a_2(z), ...a_M(z)] \\ K_2(x, y) &= b(x)^T b(y); & b(z) &= [b_1(z), b_2(z), ...b_M(z)] \end{aligned}$$

Right expression in the question can be evaluated like following:

$$\begin{aligned} & K_1(x, y)K_2(x, y) \\ &= (\sum_{m=1}^{M} a_m(x)a_m(y))(\sum_{n=1}^{N} b_n(x)b_n(y)) \\ &= \sum_{m=1}^{M} \sum_{n=1}^{N} [a_m(x)b_n(x)][a_m(y)b_n(y)] \\ &= \sum_{m=1}^{M} \sum_{n=1}^{N} c_{mn}(x)c_{mn}(x) \end{aligned}$$

where $c(z)$ is $M \times N$ dimensional vector such that $c_{mn}(z) = a_m(z)b_n(z)$
Since $K_4(x, y)$ can be written as inner product using a feature map c, using Mercer's theorem, we can say that $K_4(x, y)$ is a valid kernel.

(c) (1 point) $K_5(x, y) = (x^T y + 1)^{73}$

Solution: $K_5(x, y)$ is a valid kernel.
Proof: $x^T y$ is a valid kernel using the definition of kernels. Now adding positive constant to valid kernel results in a valid kernel (as proved in part-a), hence $x^T y + 1$ is a valid kernel. Also, we have proved in part-b that product of valid kernels results in a valid kernel, hence $(x^T y + 1) \times (x^T y + 1) \times \ .... \ \times (x^T y + 1)$ will also result in a valid kernel. Therefore, $K_5(x, y)$ is a valid kernel.

(d) (1 point) $K_6(x, y) = 6K_1(x, y) - 3K_2(x, y)$

Solution: K_6 is not a valid kernel.
Reason:

$$\begin{aligned} & X^T K_6 X \\ &= X^T (6K_1 - 3K_2) X \\ &= X^T (6K_1) X - X^T (3K_2) X \\ &= 6(X^T K_1 X) - 3(X^T K_2 X) \\ &= \quad \geqslant 0 \quad \geqslant 0 \end{aligned}$$

Since subtraction of 2 positive numbers can result in a negative and above expression can be negative for some X. Hence, K_6 is not positive semi-definite and therefor we can conclude that K_6 is not a valid kernel.

(e) (1 point) $K(x, y) = \exp(2x^T y)$ (Hint: Consider polynomial expansion of $\exp(t)$.)

Solution: K is a valid kernel.
Proof:
Doing polynomial expansion of $\exp(z)$, we get

$$exp(z) = lim_{i->\infty}(1 + z + ... + \frac{z^i}{i!})$$

After substituting z with $2z$, we get

$$exp(2z) = lim_{i->\infty}(1 + 2z + ... + \frac{(2z)^i}{i!})$$

In above equation, we can substitute $z = x^T y$ to get K.
We have proved following statements in above parts:
1) Linear combinations of valid kernels (using a positive multiplier) results in a valid kernel (proved in $part - a$)
2) Products of multiple valid kernels also result in a valid kernel (proved in $part - b$)
Now, based on above statements and the polynomial expansion of $\exp(z)$, we can conclude that $exp(K(x, y))$ will be a valid kernel.

5. (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.

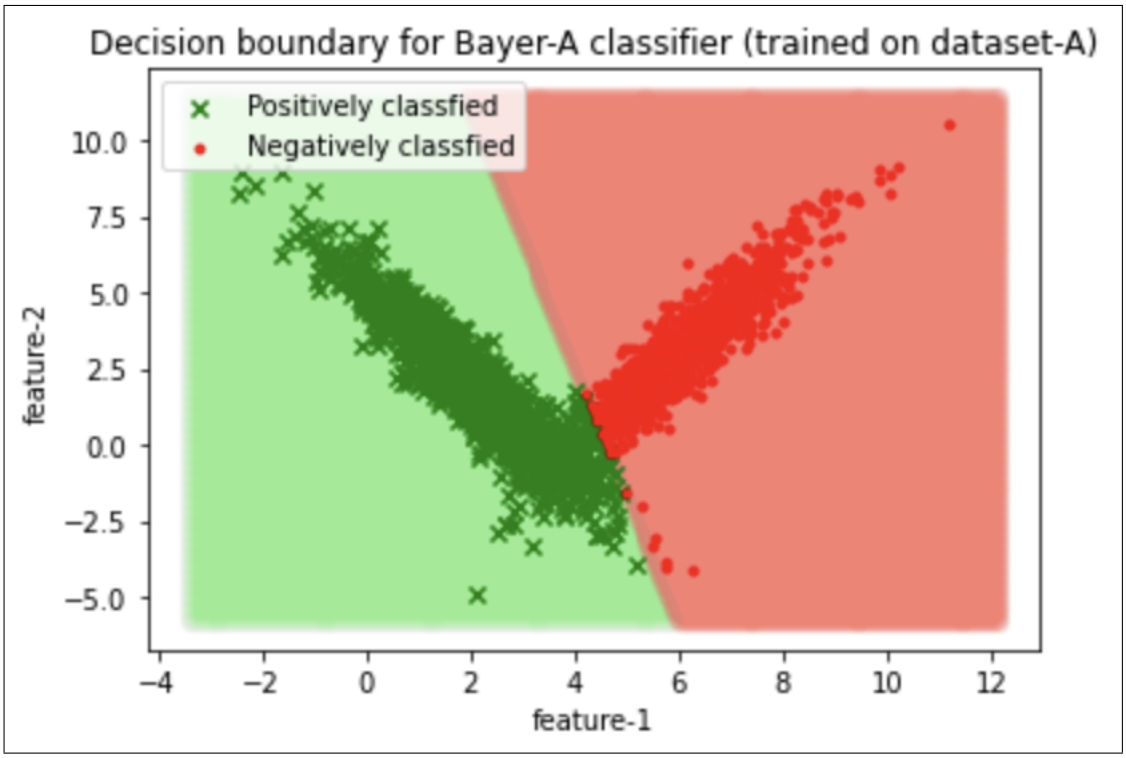
   1. **BayesA** Assume $X|Y = -1 \sim \mathcal{N}(\mu_-, I)$ and $X|Y = 1 \sim \mathcal{N}(\mu_+, I)$
   2. **BayesB** Assume $X|Y = -1 \sim \mathcal{N}(\mu_-, \Sigma)$ and $X|Y = 1 \sim \mathcal{N}(\mu_+, \Sigma)$
   3. **BayesC** Assume $X|Y = -1 \sim \mathcal{N}(\mu_-, \Sigma_-)$ and $X|Y = 1 \sim \mathcal{N}(\mu_+,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

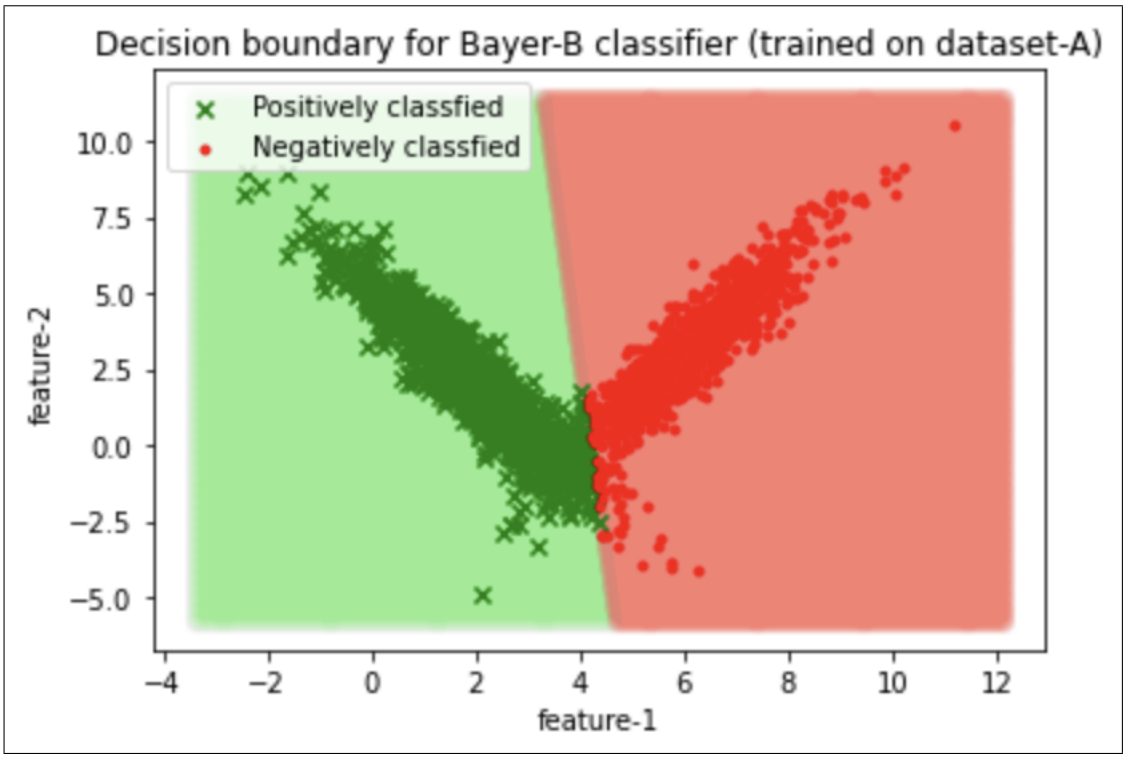
   Please see this folder for the template .ipynb file containing the helper functions, and you've to add the missing code to this file (specifically, three functions function_for_A, function_for_B, function_for_C and associated plotting/ROC code snippets) to implement the above three algorithms for the 2 datasets given in the same folder.

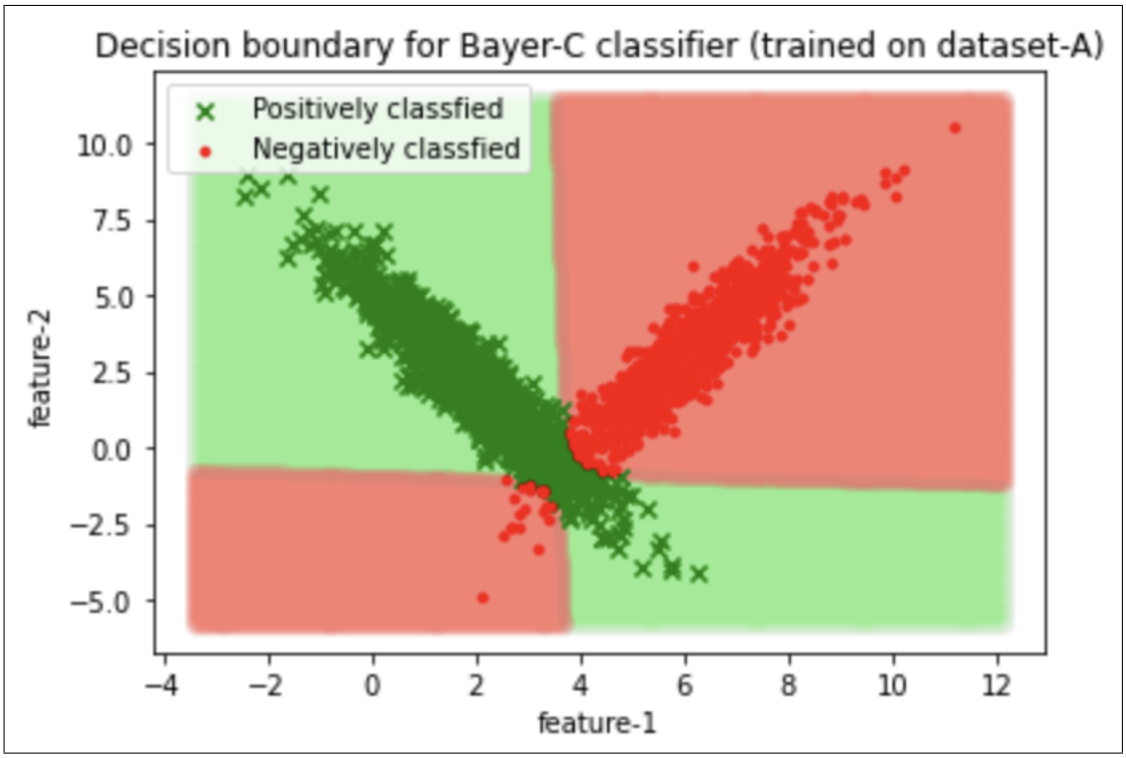
   (Note: Please provide your results/answers in the pdf file you upload to GradeScope, but submit your code separately in this moodle link. The code submitted should be a rollno1_rollno2.zip file containing a folder named Q5 with two files: rollno1_rollno2.ipynb file (including your code as well as the exact same results/plots uploaded to Gradescope) and the associated rollno1_rollno2.py file.)

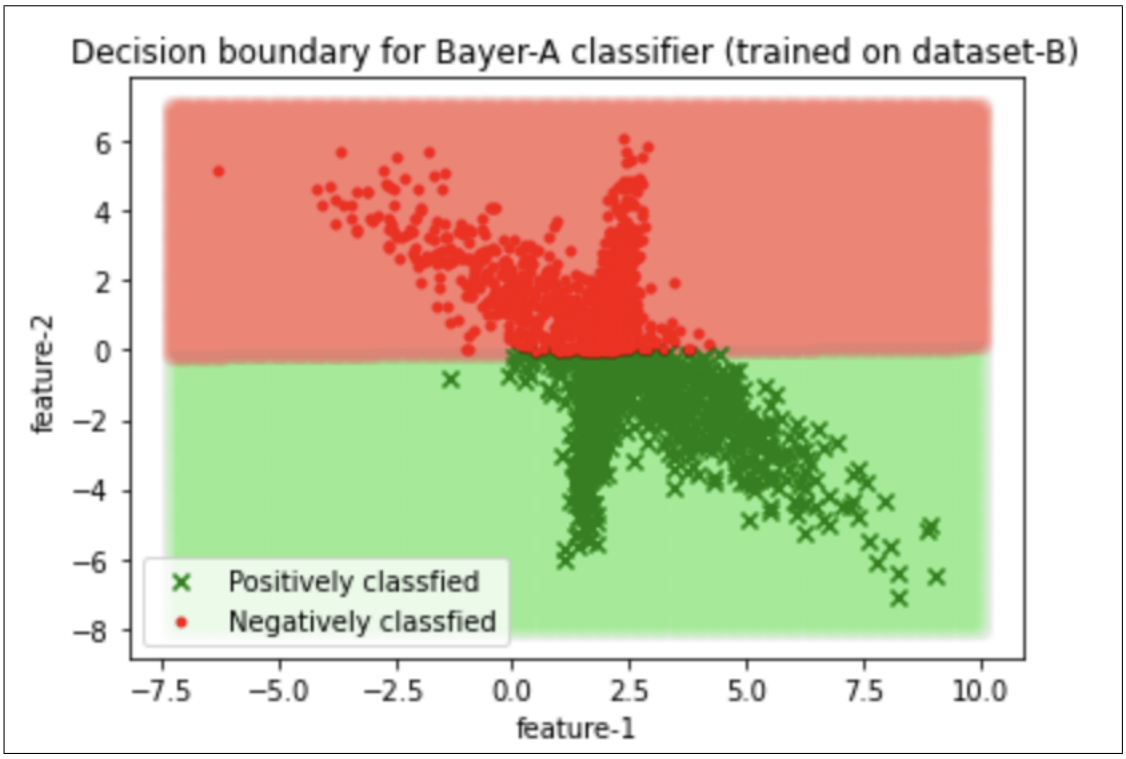
   (a) (3 points) Plot all the classifiers (3 classification algorithms on 2 datasets = 6 plots) on a 2D plot, Add the training data points also on the plots. (Color the positively classified area light green, and negatively classified area light red as in Fig 4.5 in Bishop's book).

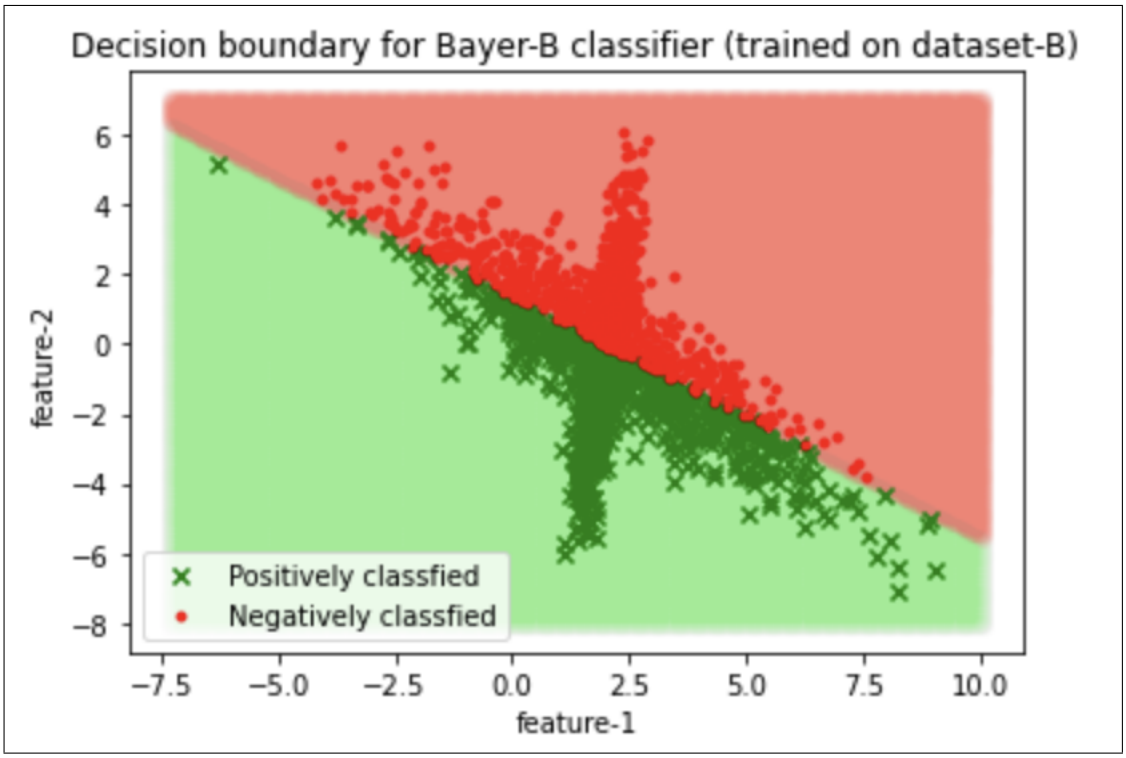
   Solution:

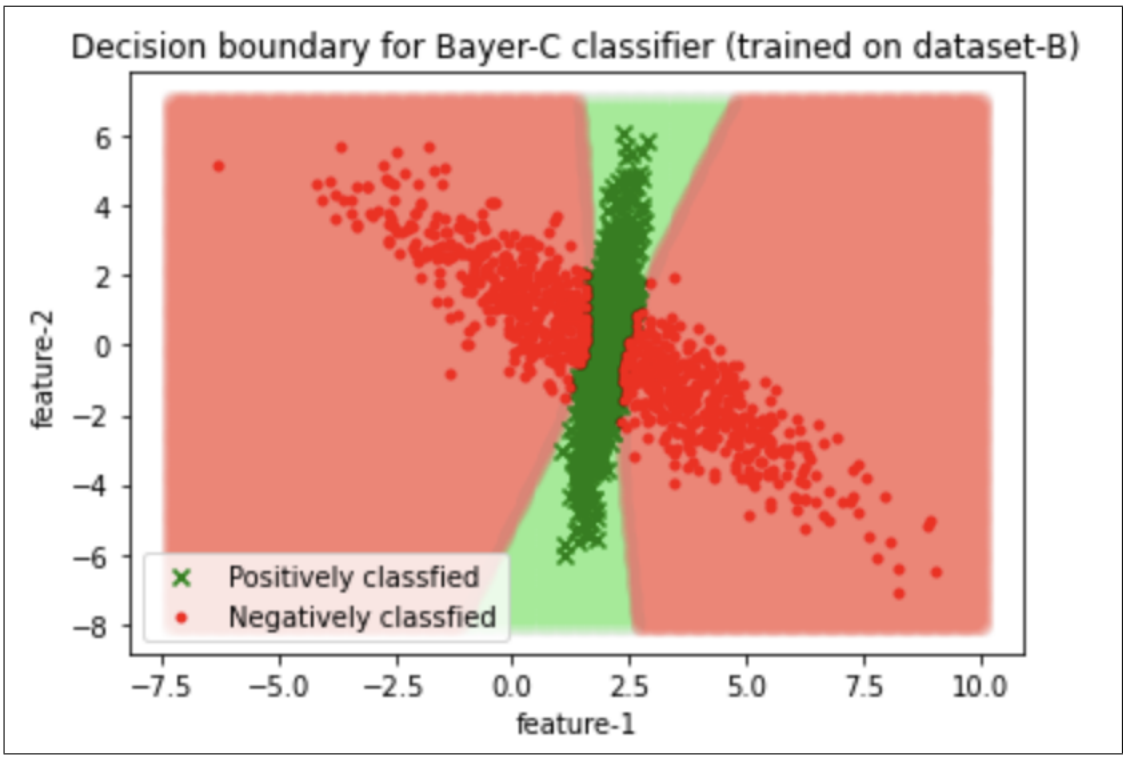

Decision boundary for Bayer-A classifier (trained on dataset-A)
Positively classfied
Negatively classfied
feature-2
feature-1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−4
−2
0
2
4
6
8
10
12


Decision boundary for Bayer-B classifier (trained on dataset-A)
Positively classfied
Negatively classfied
feature-2
feature-1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−4
−2
0
2
4
6
8
10
12


Decision boundary for Bayer-C classifier (trained on dataset-A)
Positively classfied
Negatively classfied
feature-2
feature-1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−4
−2
0
2
4
6
8
10
12

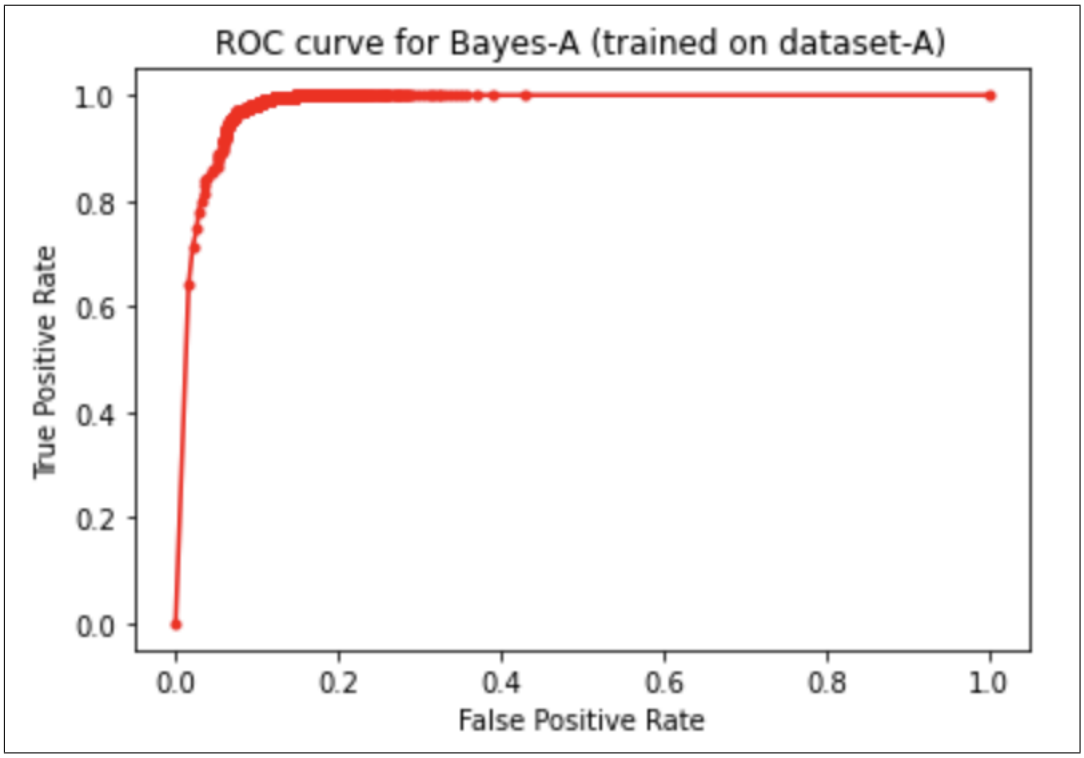

Decision boundary for Bayer-A classifier (trained on dataset-B)
Positively classfied
Negatively classfied
feature-2
feature-1
6
4
2
0
−2
−4
−6
−8
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

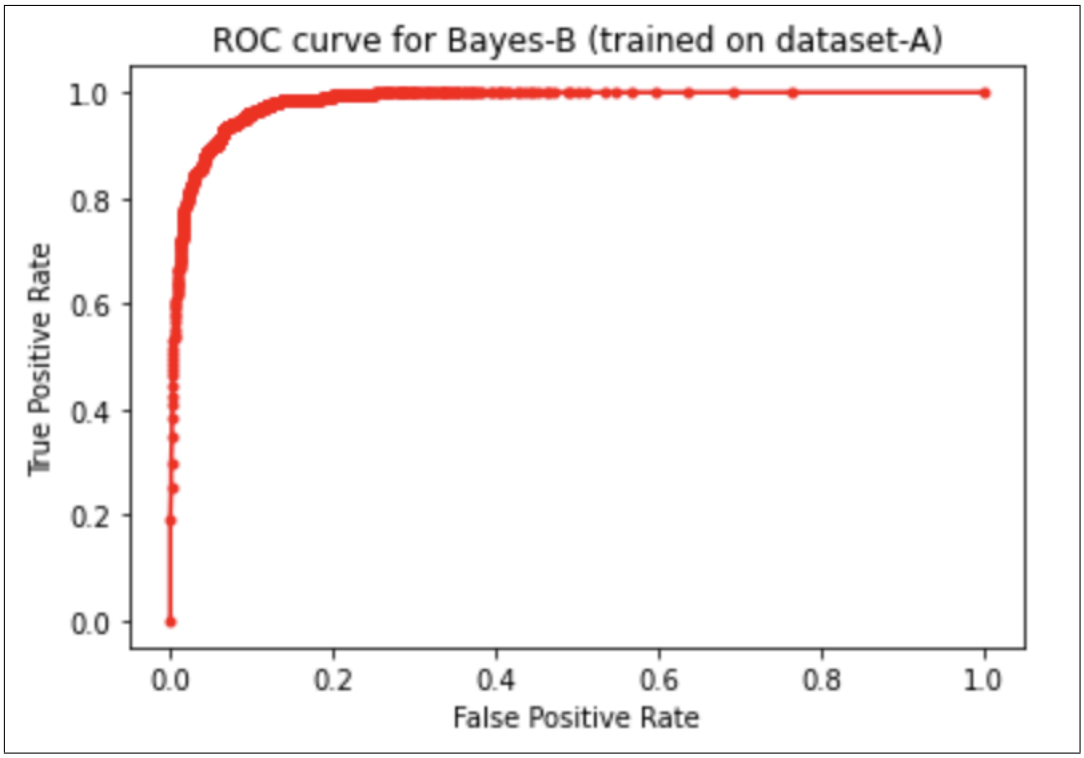

Decision boundary for Bayer-B classifier (trained on dataset-B)
feature-2
feature-1
Positively classfied
Negatively classfied
6
4
2
0
−2
−4
−6
−8
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

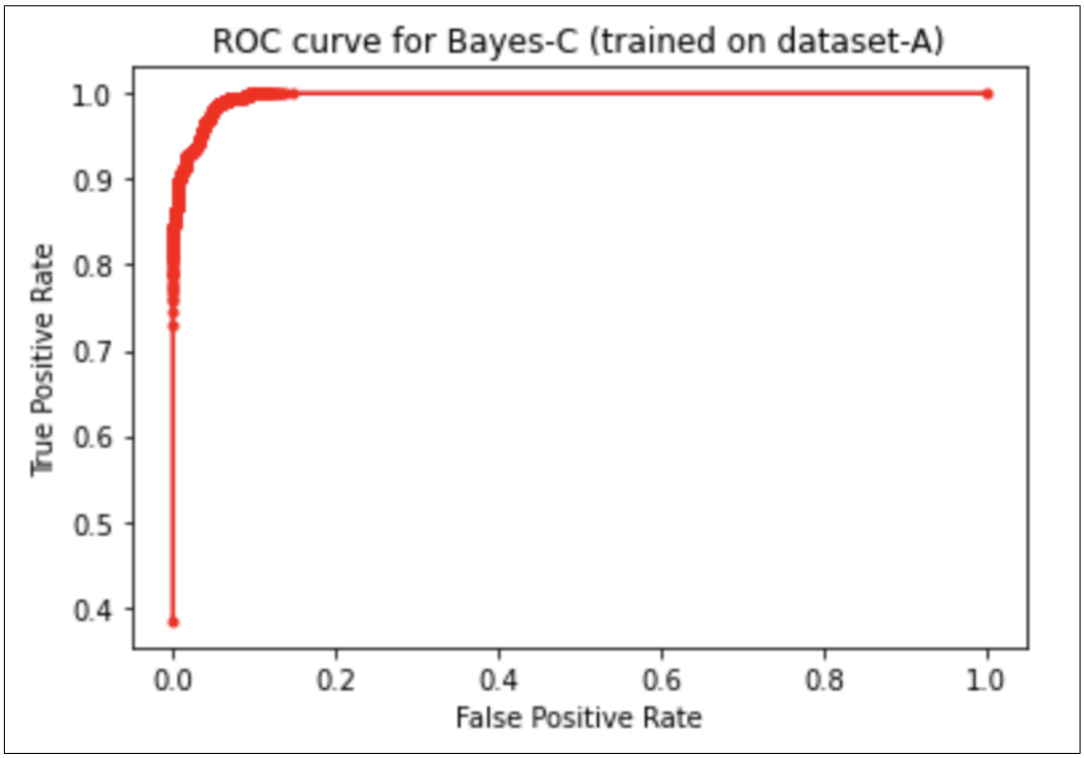

Decision boundary for Bayer-C classifier (trained on dataset-B)
feature-2
feature-1
Positively classfied
Negatively classfied
6
4
2
0
−2
−4
−6
−8
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

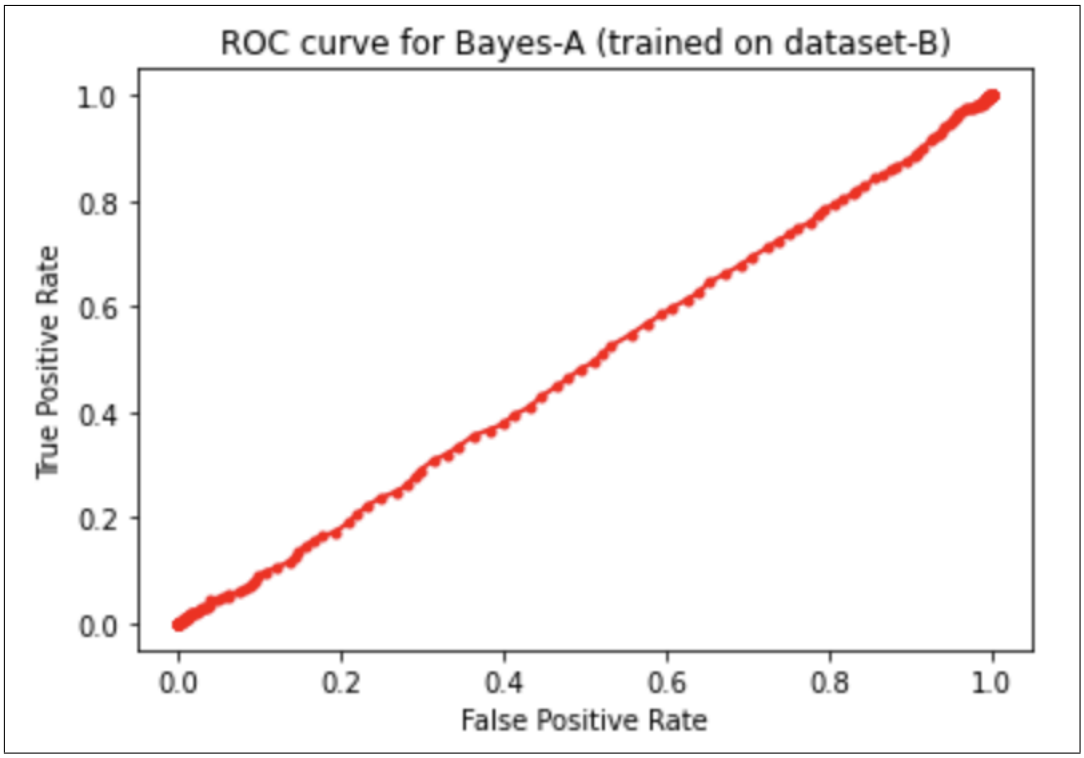
(b) (3 points) Give the ROC curves for all the classifiers. Note that a ROC curve plots the FPR (False Positive Rate) on the x-axis and TPR (True Positive Rate) on the y-axis.

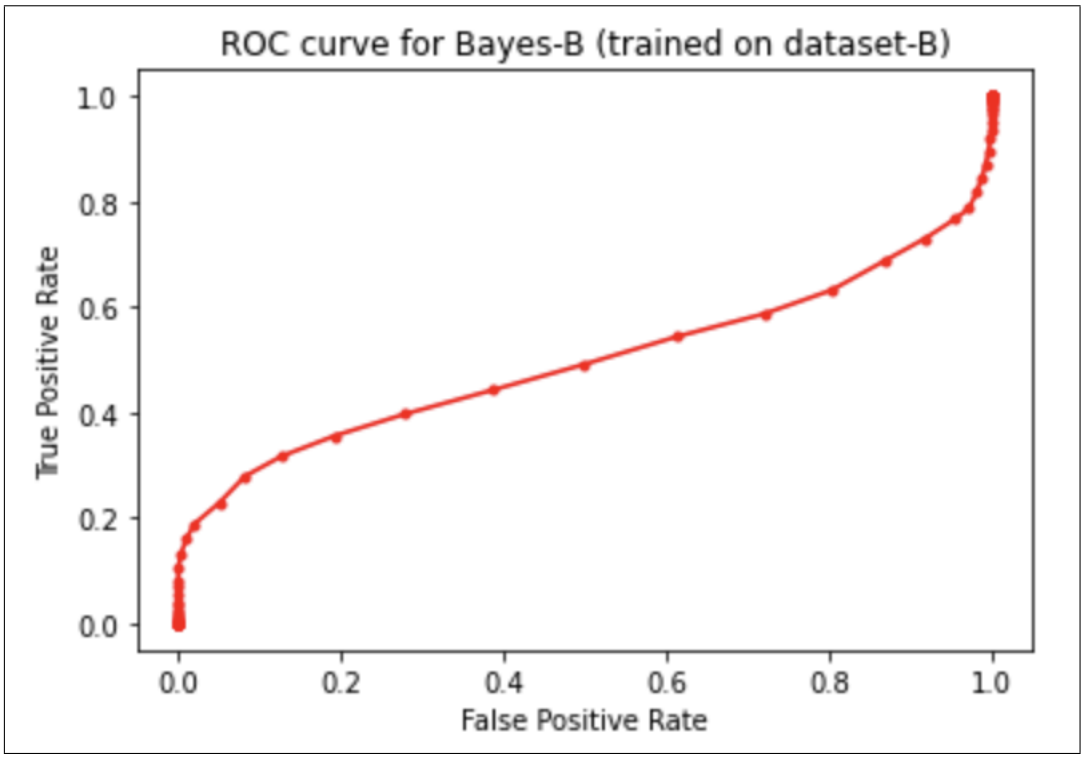
Solution:

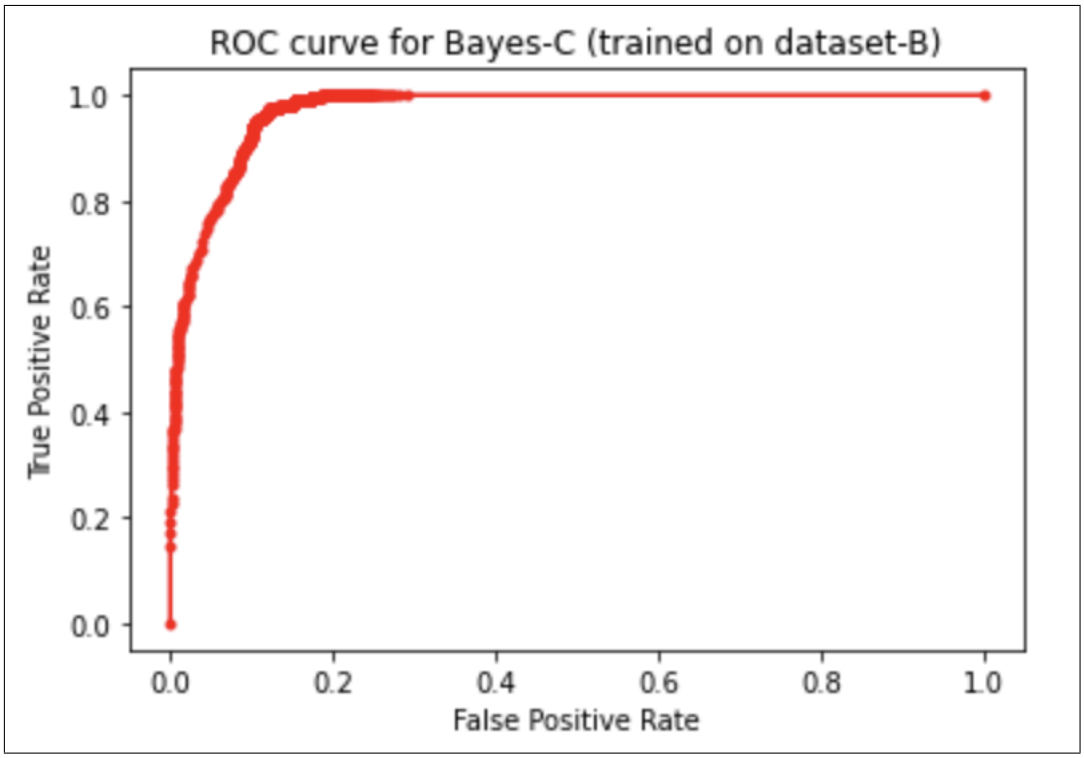



ROC curve for Bayes-B (trained on dataset-A)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve for Bayes-C (trained on dataset-A)
True Positive Rate
False Positive Rate
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve for Bayes-A (trained on dataset-B)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve for Bayes-B (trained on dataset-B)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



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

Solution:

	Dataset-A	**Dataset-B**
Bayes-A	0.066	0.5085
Bayes-B	0.0675	0.504
Bayes-C	0.034	0.083

(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:
Assumptions

We have assumed that the conditional distribution of data belongs to the gaussian family for all the classifiers.

Additionally,

1. For Bayes-A, we have assumed that features are independent of each other when conditioned over classes.
2. For Bayes-B, we have assumed that conditional distribution of data share the same covariance matrix for all the classes.

Observations

The decision boundary for Bayes-A, Bayes-B is linear (as shown in part-a), and hence both models are linear classifiers. Bayes-C is a complex model as it's decision boundary is complex as expected.

Dataset-B is more complex as compared to dataset-A and simple linear models are performing very poorly on dataset-B. Bayes-C performed decently on dataset-B. We can conclude that linear models like Bayes-A, Bayes-B perform poorly, especially when two classes cannot be separated using a linear boundary and data (dataset-B in our case) does not meet the assumptions we were making before training the model. Therefore, we are getting significantly less error value for the Bayes-C model (as we are not making many assumptions about the data for Bayes-C) while a large value for Bayes-A, Bayes-B.

Note on non-parametric models

We can use non-parametric models like the KMeans classifier for solving this problem instead of the Bayes classifier. The significant advantage of using them is that we need not make any assumption on the distribution family. Hence, non-parametric models will work even if data does not belong to some specific distribution.

However, the main limitation of using non-parametric models over parametric models is that we cannot scale them on massive datasets. The complexity of these models increases as dataset size increases which in contrast remains fixed for the parametric models.

6. (5 points) [CODING A DIFFERENT DENSITY ESTIMATION?] In the previous question, the class conditional densities were Gaussian. But not all real-world datasets are Gaussian as is to begin with. For instance, consider this data on expression/activity level of genes in the skeletal muscle tissue of different individuals, provided as a "Genes $\times$ Samples" matrix in this link.
(Note: Put all your code pertaining to this question into a single file rollno1_rollno2_**genes**.<fileextension>, and include this single file inside the Q6 folder of the rollno1_rollno2.zip file mentioned in the previous question.)

   (a) (2 points) (Model Selection) How would you model any given gene in this dataset, i.e., what distribution will you assume for a gene? Assume that every gene follows the same parametric model/distribution, but with different parameter values. Support your assumption.

Solution: Since data is sparse and there are very few unique values for each feature in the data, we can possibly assume data distribution to be **multi-categorical**.

(b) (2 points) (MLE Code) How will you obtain the MLE estimates of the assumed model's parameters? (no need to derive it, just state your answer as a closed-form formula or as an optimization method). Write a code to estimate these parameters for each gene.

Solution:

(c) (1 point) (Diagnostic Plots) Use your code to also plot the sample mean (x-axis) vs. sample variance (y-axis) of each gene (across all genes, with each dot in this scatter-plot being a gene). Overlay on this plot using a different color, the model mean vs. variance of each gene (i.e., mean/variance calculated using the expectation/variance formula implied by the model/distribution learnt via MLE). What does this plot tell you?

Solution: