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

Release Date: Sep 28, 2021
Due Date : Oct 7, 23:59

Roll No: ME18B030 Name: S Jeeva
Roll No: ME18B046 Name: G Deepak
Team number: 8
References (if any): Geogebra, StackOverflow, CM Bishop, CS5691 lecture slides

- This assignment has to be completed in teams of two. Collaborations outside the team are strictly prohibited.
- 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 **N8Z67W**)
- For the programming questions, please submit your code directly in moodle (carefully following the filename/folder/README conventions given in the questions/moodle), but provide your results/answers in the pdf file you upload to GradeScope. We will run plagiarism checks on codes, and any detected plagiarism in writing/code will be strictly penalized.
- 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).
- Check the Moodle discussion/announcement forums regularly for updates regarding the assignment. Please start early and clear all doubts ASAP. Post your doubt only on Moodle Discussion Forum so that everyone is on the same page. Please note that the TAs can **only** clarify doubts regarding problem statements (they won't discuss any prospective solution or verify your solution or give hints).

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:

(a) For an exponential distribution we have,

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

where λ is the rate parameter and μ is the mean of the distribution.

$$\mu = E[X] = \int_{-\infty}^{\infty} x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

(i) Maximum Likelihood Estimation :

$\rightarrow$ Let $L \equiv L(\theta; D_N)$ denote the likelihood function

$$L(\theta; D_N) = p(\{x1, \ldots, x6\}/\lambda)$$

where $\{x_1 = 2, x_2 = 6, x_3 = 7, x_4 = 1, x_5 = 4, x_6 = 3\} \in D_N$

$$\implies L(\theta; D_N) = \prod_{i=1}^{6} p_{X/\lambda}(x_i/\lambda)$$

$$\implies L(\theta; D_N) = \lambda^6 e^{-\lambda \sum_{i=1}^{6} x_i}$$

$\rightarrow$ Let $LL(\theta; D_N)$ represent the log-likelihood function

$$LL(\theta; D_N) = 6log\lambda - \lambda \sum_{i=1}^{6} x_i$$

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} \; LL(\theta; D_N)$$

$$\implies \hat{\lambda}_{ML} = \frac{6}{\sum_{i=1}^{6} x_i} = \frac{6}{23}$$

$$\hat{\mu}_{ML} = \frac{1}{\hat{\lambda}_{ML}}$$

$$\implies \hat{\mu}_{ML} = \sum_{i=1}^{6} x_i = \frac{23}{6}$$

(ii) Bias of $\hat{\mu}_{ML}$:

$$Bias(\hat{\mu}_{ML}) = E_{D_N}[\hat{\mu}_{ML}] - \mu$$

$$\text{Bias}(\hat{\mu}_{ML}) = E_{D_N}\left[\frac{\sum_{i=1}^{6} x_i}{6}\right] - \mu$$

$$E_{D_N}\left[\frac{\sum_{i=1}^{6} x_i}{6}\right] = \frac{1}{6}\sum_{i=1}^{6} E[x_i] = E[X] = \mu$$

$$\implies \text{Bias}(\hat{\mu}_{ML}) = \mu - \mu = 0$$

Bias of $\hat{\lambda}_{ML}$:

$$\text{Bias}(\hat{\lambda}_{ML}) = E_{D_N}[\hat{\lambda}_{ML}] - \lambda$$

$$\text{Bias}(\hat{\lambda}_{ML}) = E_{D_N}\left[\frac{6}{\sum_{i=1}^{6} x_i}\right] - \lambda$$

From Gamma distribution we can get,

$$E_{D_N}\left[\frac{N}{\sum_{i=1}^{N} x_i}\right] = NE_{D_N}\left[\frac{1}{\sum_{i=1}^{N} x_i}\right] = N\frac{\lambda}{N-1}$$

$$\implies \text{Bias}(\hat{\lambda}_{ML}) = 6\frac{\lambda}{6-1} - \lambda = \frac{\lambda}{5}$$

(iii) Unbiased Estimator :

We only have the estimator $\hat{\lambda}_{ML}$ to be biased with a bias of $\frac{\lambda}{5}$.

In general for any given 'N' we can set an unbiased estimate for $\hat{\lambda}_{unb}$ as,

$$\hat{\lambda}_{unb} = \frac{N-1}{\sum_{i=1}^{N} x_i}$$

$$\therefore \text{Bias}(\hat{\lambda}_{unb}) = \frac{(N-1)\lambda}{N-1} - \lambda = 0$$

Thus, we get an unbiased estimate for μ as $\hat{\mu}_{ML}$ and for λ as $\hat{\lambda}_{unb}$.

(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. (1½ 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:

(b) (i) Given prior distribution on θ as Pareto distribution.

$$\theta \sim Pareto(\alpha, \beta)$$

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

Also, we have data $D = (x_1, x_2, \ldots, x_n)$ where each $X_i \sim Uniform(0, \theta)$ and are i.i.d

$$Posterior = p(\theta/D) = \frac{p(D/\theta)p(\theta)}{p(D)}$$

where $p(D/\theta)$ is the likelihood, $p(\theta)$ is the prior on θ and $p(D)$ is the normalizing constant that represents evidence.

$$p(D/\theta) = p(\{x_1, \ldots, x_n\}/\theta) = \prod_{i=1}^{N} p(x_i/\theta) = \prod_{i=1}^{N} \frac{1}{\theta} = \theta^{-N}$$

$$p(D) = \int_{-\infty}^{\infty} p(D/\theta)p(\theta)d\theta$$

Hence we can write the expression for Posterior as,

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

$$p(D) = \int_{-\infty}^{\infty} p(D/\theta)p(\theta)d\theta$$

$$\implies p(D) = \int_{-\infty}^{\infty} \alpha\beta^{\alpha}\theta^{-\alpha-N-1}\mathbb{1}_{(\beta,\infty)}(\theta)d\theta = \int_{\beta}^{\infty} \alpha\beta^{\alpha}\theta^{-\alpha-N-1}(\theta)d\theta$$

$$\implies p(D) = \alpha\beta^{\alpha}\frac{\theta^{-\alpha-N}}{-\alpha-N}\Big|_{\beta}^{\infty} = \frac{\alpha\beta^{\alpha}}{N+\alpha}\beta^{-\alpha-N}$$

Thus we can calculate the posterior probability as,

$$p(\theta/D) = \frac{\alpha\beta^{\alpha}\theta^{-\alpha-N-1}\mathbb{1}_{(\beta,\infty)}(\theta)}{\frac{\alpha\beta^{\alpha}}{N+\alpha}\beta^{-\alpha-N}}$$

$$\implies p(\theta/D) = (\alpha+N)\beta^{\alpha+N}\theta^{-(\alpha+N)-1}\mathbb{1}_{(\beta,\infty)}$$

The above posterior distribution is also a Pareto distribution with $\alpha^{'=N+\alpha}$ and $beta$

$$\implies \theta/D \sim Pareto(\alpha+N, \beta)$$

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

Solution:

(ii) We got the posterior distribution as $\theta/D \sim Pareto(N+\alpha, \beta)$

Maximum a posteriori Estimate (MAP) :

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \; p(\theta/D)$$

$$\therefore \hat{\theta}_{MAP} = \lim_{\theta \to \beta^+} \theta$$

This is because $p(\theta/D) \sim Pareto(N+\alpha, \beta)$ is maximum as $\theta \to \beta^+$

Maximum Likelihood Estimate (MLE) :

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \; p(D/\theta)$$

We know $\beta \leqslant \theta < \infty$ from the Pareto prior distribution of $\theta \sim Pareto(\alpha, \beta)$

$$\therefore \hat{\theta}_{MLE} = \lim_{\theta \to \beta^+} \theta$$

$$\implies \hat{\theta}_{MAP} = \hat{\theta}_{MLE} = \lim_{\theta \to \beta^+} \theta$$

This is because $p(D/\theta) = \theta^{-N}$ is maximum as $\theta \to \beta^+$, as $\beta < \theta < \infty$ Thus, we get both the MAP estimate ($\hat{\theta}_{MAP}$) and ML estimate ($\hat{\theta}_{MLE}$) to be equal to $\lim_{\theta \to \beta^+} \theta$.

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:

(iii) The posterior distribution can be given as $\theta/D \sim Pareto(N+\alpha, \beta)$

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

$$E_D[L] = E[L(\theta, \hat{\theta})/D] = \int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 p(\theta/D) d\theta$$

$$E_D[L] = \int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 (N+\alpha) \beta^{N+\alpha} \theta^{-N-\alpha-1} \mathbb{1}_{(\beta,\infty)} d\theta$$

$$E_D[L] = \int_\beta^\infty (\theta - \hat{\theta})^2 (N+\alpha)\beta^{N+\alpha}\theta^{-N-\alpha-1} d\theta$$

$$E_D[L] = (N+\alpha)\beta^{N+\alpha}\left[\frac{\theta^{-N-\alpha+2}}{-N-\alpha+2} - 2\hat{\theta}\frac{\theta^{-N-\alpha+1}}{-N-\alpha+1} + \hat{\theta}^2\frac{\theta^{-N-\alpha}}{-N-\alpha}\right]\Big|_\beta^\infty$$

$$\implies E_D[L] = \frac{N+\alpha}{\beta^{-(N+\alpha)}}\left[\frac{\beta^{-N-\alpha+2}}{N+\alpha-2} - 2\hat{\theta}\frac{\beta^{-N-\alpha+1}}{N+\alpha-1} + \hat{\theta}^2\frac{\beta^{-N-\alpha}}{N+\alpha}\right]$$

$$\implies E_D[L] = (N+\alpha)\left[\frac{\beta^2}{N+\alpha-2} - 2\hat{\theta}\frac{\beta}{N+\alpha-1} + \hat{\theta}^2\frac{1}{N+\alpha}\right]$$

$E_D[L]$ is a function of $\hat{\theta}$. Since it is a quadratic function with co-efficient of highest degree term being positive, we get a minima for the function.

$$\frac{d(E_D[L])}{d\hat{\theta}} = 0$$

So we get,

$$-\frac{2\beta}{N+\alpha-1} + \frac{2\hat{\theta}_{min}}{N+\alpha} = 0$$

$$\implies \hat{\theta}_{min} = \beta\left(\frac{N+\alpha}{N+\alpha-1}\right)$$

The estimator $\hat{\theta}_{min} = \beta\left(\frac{N+\alpha}{N+\alpha-1}\right)$ minimizes the posterior expected square loss.

We can also observe that $\hat{\theta}_{min}$ is the same as $E_D[\theta]$ which is same as $E[\theta/D]$

$$E[\theta/D] = \int_{-\infty}^{\infty} \theta p(\theta/D) d\theta$$

$$E[\theta/D] = \int_{-\infty}^{\infty} \theta(N+\alpha)\beta^{N+\alpha}\theta^{-N-\alpha-1}\mathbb{1}_{(\beta,\infty)} d\theta$$

$$E[\theta/D] = \int_\beta^\infty (N+\alpha)\beta^{N+\alpha}\theta^{-N-\alpha} d\theta$$

$$E[\theta/D] = \left(\frac{N+\alpha}{N+\alpha-1}\right)\frac{\beta^{N+\alpha}}{\beta^{N+\alpha-1}}$$

Thus,

$$\hat{\theta}_{min} = E[\theta/D] = \beta\left(\frac{N+\alpha}{N+\alpha-1}\right)$$

(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:

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

(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:

(c) (i) We can get the probability estimate from Parzen window estimator as,

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h} k\left(\frac{x - x_i}{h}\right) = \frac{1}{Nh}\sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right)$$

Since, $(x_1, \ldots, x_n)$ are i.i.d samples drawn from $U(0,1)$ we can write,

$$E_D[\hat{p}(x)] = E_D\left[\frac{1}{Nh}\sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right)\right] = \frac{1}{Nh}\sum_{i=1}^{N} E_D\left[k\left(\frac{x - x_i}{h}\right)\right]$$

$$E_D[\hat{p}(x)] = \frac{N}{Nh} E_D\left[k\left(\frac{x - x_1}{h}\right)\right] = \frac{1}{h} E_D\left[k\left(\frac{x - x_1}{h}\right)\right]$$

Since we know that $0 \leqslant x_1 \leqslant 1$ because $x_1 \sim U(0,1)$,

$$E_D[\hat{p}(x)] = \frac{1}{h}\int_{-\infty}^{\infty} k\left(\frac{x - x_1}{h}\right) dx_1 = \frac{1}{h}\int_{0}^{1} k\left(\frac{x - x_1}{h}\right) dx_1$$

Also the kernel function $k\left(\frac{x-x_1}{h}\right)$ can be given as,

$$k\left(\frac{x-x_1}{h}\right) = \begin{cases} 0 & x - x_1 \leqslant 0 \\ e^{-\left(\frac{x-x_1}{h}\right)} & x - x_1 > 0 \end{cases}$$

Now we can re-write $E_D[\hat{p}(x)]$ as,

$$E_D[\hat{p}(x)] = \frac{1}{h}\int_0^1 k\left(\frac{x-x_1}{h}\right) dx_1$$

Case 1 : When $x < 0$ we get $(x - x_1) < 0$ which means that $k\left(\frac{x-x_1}{h}\right) = 0$

$$E_D[\hat{p}(x)] = \frac{1}{h}\int_0^1 (0) dx_1 = 0$$

Case 2 : When $0 \leqslant x \leqslant 1$,

$$\begin{cases} k\left(\frac{x-x_1}{h}\right) = e^{-\left(\frac{x-x_1}{h}\right)} & 0 \leqslant x_1 < x \\ k\left(\frac{x-x_1}{h}\right) = 0 & x \leqslant x_1 \leqslant 1 \end{cases}$$

$$E_D[\hat{p}(x)] = \frac{1}{h}\int_0^x k\left(\frac{x-x_1}{h}\right) dx_1 + 0 = \frac{1}{h}\int_0^x e^{-\left(\frac{x-x_1}{h}\right)} dx_1$$

$$E_D[\hat{p}(x)] = e^{-\left(\frac{x-x_1}{h}\right)}\Big|_0^x = 1 - e^{-\frac{x}{h}}$$

Case 3 : When $x \geqslant 1$,

$$\begin{cases} k\left(\frac{x-x_1}{h}\right) = e^{-\left(\frac{x-x_1}{h}\right)} & 1 \leqslant x_1 < x \\ k\left(\frac{x-x_1}{h}\right) = 0 & x \leqslant x_1 \leqslant \infty \end{cases}$$

$$E_D[\hat{p}(x)] = \frac{1}{h}\int_1^x k\left(\frac{x-x_1}{h}\right) dx_1 + 0 = \frac{1}{h}\int_0^x e^{-\left(\frac{x-x_1}{h}\right)} dx_1$$

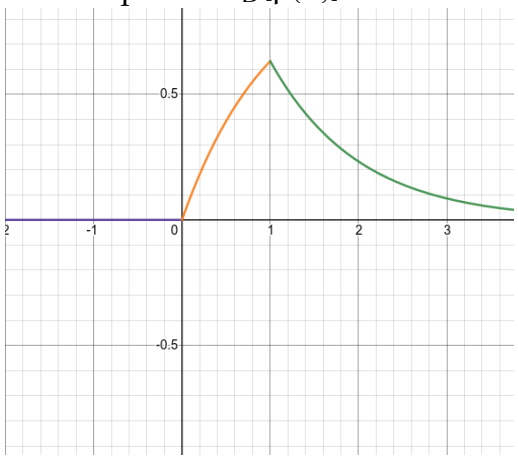
$$E_D[\hat{p}(x)] = e^{-\left(\frac{x-x_1}{h}\right)}\Big|_1^x = e^{-\frac{x-1}{h}} - e^{-\frac{x}{h}} = e^{\frac{1-x}{h}} - e^{-\frac{x}{h}}$$

Hence we get,

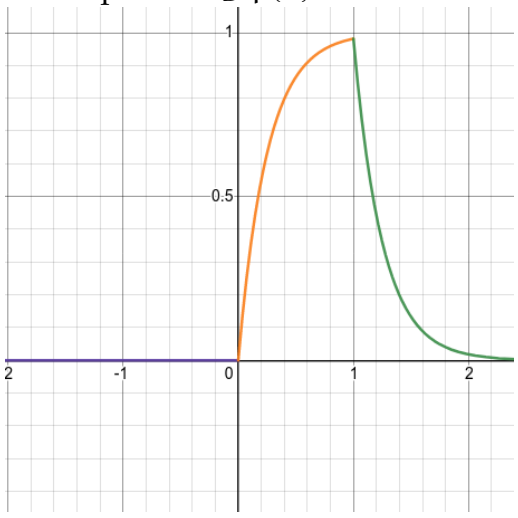
$$E_D[\hat{p}(x)] = \begin{cases} 0 & -\infty < x < 0 \\ 1 - e^{-\frac{x}{h}} & 0 \leqslant x \leqslant 1 \\ e^{\frac{1-x}{h}} - e^{-\frac{x}{h}} & 1 \leqslant x < \infty \end{cases}$$

(ii)

The plot of $E_D[\hat{p}(x)]$ for h = 1



The plot of $E_D[\hat{p}(x)]$ for h = 0.25



The plot of $E_D[\hat{p}(x)]$ for $h = 0.0625$

We observe that the expectation of the parzen window estimator converges closer and closer to the $U(0, 1)$ as the value of smoothing parameter h decreases from 1 to 0.0625.

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:

(a) Since, the objective is to find the optimal Bayes classifier $h(x)$ for the given data and the loss matrix it is logical to asssume that the data is normally distributed with different means and shared covariance when conditioned under each of the three classes.

$$X/Y = 1 \sim N(\mu_1, \sigma_s)$$

$$X/Y = 2 \sim N(\mu_2, \sigma_s)$$

$$X/Y = 3 \sim N(\mu_3, \sigma_s)$$

We can also find the prior probabilities for classes 1,2, and 3 using the number of data points,

$$p(Y = 1) = \frac{N_1}{N} = \frac{5}{14}, p(Y = 2) = \frac{N_2}{N} = \frac{4}{14}, p(Y = 3) = \frac{N_3}{N} = \frac{5}{14}$$

where $N_1 = 5, N_2 = 4, N_3 = 5, N = N_1 + N_2 + N_3 = 14$

Maximum Likelihood Estimates :

For this Bayes Classifier with K = 3 classes, we need to estimate 4 parameters, namely, $\{\mu_1, \mu_2, \mu_3, \sigma_s\}$

$$\mu_1 = \mu_1^{ML} = \frac{1}{N_1} \sum_{i=1}^{N} \mathbb{1}_{(y_i=1)} x_i = \frac{(-2.9 - 0.7 - 2.4 - 1.4 - 3.4)}{5} = -2.16$$

$$\sigma_1 = \frac{1}{N_1} \sum_{N \in Y=1} (x_i - \mu_1)^2 = \frac{\left((-0.74)^2 + (1.46)^2 + (-0.24)^2 + (0.76)^2 + (-1.24)^2\right)}{5} = 0.9704$$

$$\mu_2 = \mu_2^{ML} = \frac{1}{N_2} \sum_{i=1}^{N} \mathbb{1}_{(y_i=2)} x_i = \frac{(0.4 - 0.3 + 0.8 + 1.2)}{4} = 0.525$$

$$\sigma_2 = \frac{1}{N_2} \sum_{N \in Y=2} (x_i - \mu_2)^2 = \frac{\left((-0.125)^2 + (-0.825)^2 + (0.275)^2 + (0.675)^2\right)}{4} = 0.3068$$

$$\mu_3 = \mu_3^{ML} = \frac{1}{N_3} \sum_{i=1}^{N} \mathbb{1}_{(y_i=3)} x_i = \frac{(1.4 + 0.9 + 1.8 + 2.3 + 2.8)}{5} = 1.84$$

$$\sigma_1 = \frac{1}{N_3} \sum_{N \in Y=3} (x_i - \mu_3)^2 = \frac{\left((-0.44)^2 + (-0.94)^2 + (-0.04)^2 + (0.46)^2 + (0.96)^2\right)}{5} = 0.4424$$

$$\sigma_s = \frac{N_1}{N}\sigma_1 + \frac{N_2}{N}\sigma_2 + \frac{N_3}{N}\sigma_3 = \frac{5}{14}(0.9704) + \frac{4}{14}(0.3068) + \frac{5}{14}(0.4424) = 0.5922$$

where,

$\{\mu_1, \mu_2, \mu_3\}$ are the means of the class conditionals for the classes 1,2 and 3 respectively.

$\{\sigma_s\}$ is the shared covariance among the three classes 1,2 and 3.

Posterior Probabilities

The posterior probabilities for the three classes w.r.t given data are,

$$p(Y=1/X) = \frac{p(X/Y=1)p(Y=1)}{\sum_{i=1}^{3} p(X/Y=i)p(Y=i)} = e^{\frac{(x-\mu_1)^2}{2\sigma_s^2}} \frac{5}{a}$$

$$p(Y=2/X) = \frac{p(X/Y=2)p(Y=2)}{\sum_{i=1}^{3} p(X/Y=i)p(Y=i)} = e^{\frac{(x-\mu_2)^2}{2\sigma_s^2}} \frac{4}{a}$$

$$p(Y=3/X) = \frac{p(X/Y=3)p(Y=3)}{\sum_{i=1}^{3} p(X/Y=i)p(Y=i)} = e^{\frac{(x-\mu_3)^2}{2\sigma_s^2}} \frac{5}{a}$$

where, $a = 14(\sqrt{2\pi\sigma_s^2}) \sum_{i=1}^{3} p(X/Y=i)p(Y=i)$

Expected Loss

$$E_{Y/X}[L(h(x)=j)] = E[L/X=x] == \sum_{i=1}^{3} L_{i,h(x)=j} p(Y=i/X)$$

$$h(x) = \arg\min_{j} \sum_{i=1}^{3} L_{i,h(x)=j} p(Y=i/X)$$

We have to find the decision boundaries by solving the above optimization problem.

$$E[L(h(x)=1)/X] = 0 + p(Y=2/X) + 2p(Y=3/X) = \frac{4}{a} e^{\frac{(x-0.525)^2}{(0.8375^2)}} + \frac{10}{a} e^{\frac{(x-1.84)^2}{(0.8375^2)}}$$

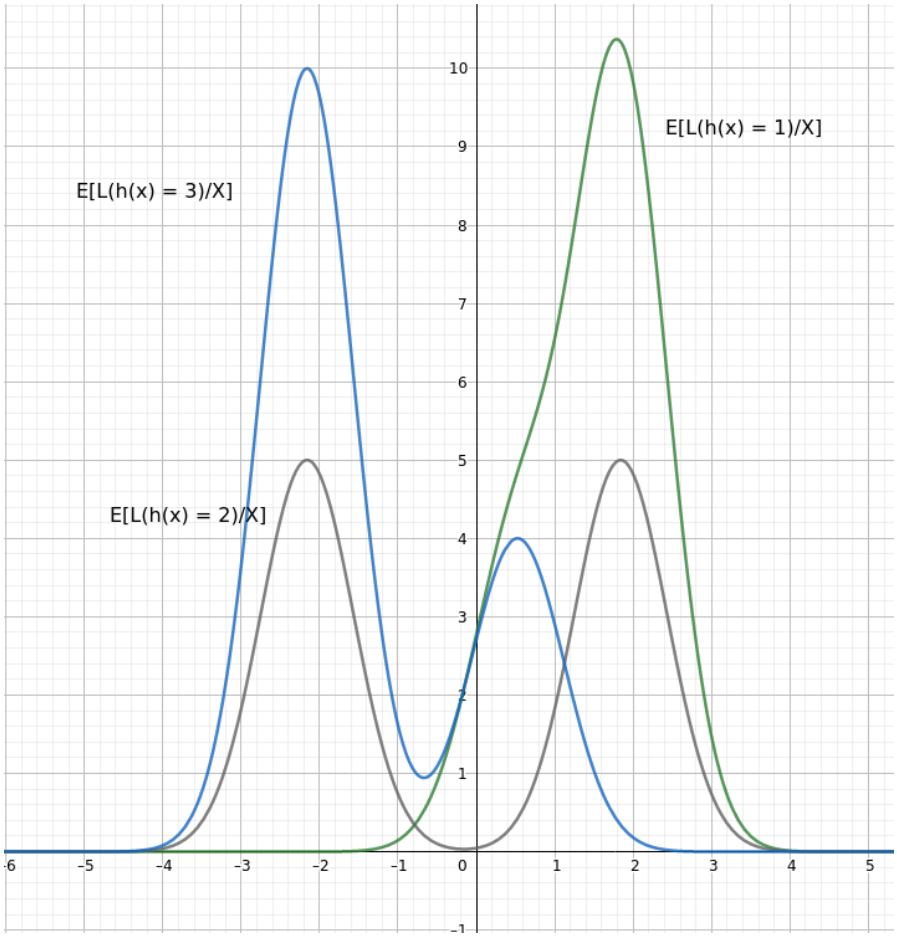
$$E[L(h(x)=2)/X] = p(Y=1/X) + p(Y=3/X) = \frac{5}{a} e^{\frac{(x+2.16)^2}{(0.8375^2)}} + \frac{5}{a} e^{\frac{(x-1.84)^2}{(0.8375^2)}}$$

$$E[L(h(x)=3)/X] = 2p(Y=1/X) + p(Y=2/X) = \frac{10}{a} e^{\frac{(x+2.16)^2}{(0.8375^2)}} + \frac{4}{a} e^{\frac{(x-0.525)^2}{(0.8375^2)}}$$

Decision Boundaries

We used "GeoGebra" to plot the expected losses $E[L(h(x)=1)/X]$, $E[L(h(x)=2)/X]$, $E[L(h(x)=3)/X]$ to find the decision boundaries because of the nature of the expected loss functions (too difficult to be solved on pen/paper). We find the minimum of the three functions $E[L(h(x)=1)/X]$, $E[L(h(x)=2)/X]$, $E[L(h(x)=3)/X]$ and thus find the decision

boundaries for the Bayes Classifier.



From the above image, we can two decision boundaries (i.e) at $x_{12}^{*} = -0.7884$ (Decision boundary b/w classes 1 and 2) and $x_{23}^{*} = 1.123$ (Decision boundary b/w classes 2 and 3)

Hence the decision for the given Bayes Classifier is,

$$h(x) = \begin{cases} 1 & -\infty < x \leqslant -0.7884 \\ 2 & -0.7884 < x \leqslant 1.123 \\ 3 & 1.123 < x < \infty \end{cases}$$

(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:
(b) (i)

$$H_1(x) = c^* = \arg\max_c\ p(c/x)$$

Calculating Risk R_1 over a single pattern x

$$R_1 = E_{x,c}[L] = E_x[E_{c/x}[L]]$$

$$\implies R_1 = E_x\left[\sum_{i=1}^{k} L_{i,H_1(x)=c^*} p(c = i/x)\right]$$

$$\implies R_1 = E_x\left[\sum_{i=1}^{k} \mathbb{1}_{i\neq H_1(x)} p(c = i/x)\right]$$

Since there's only a single pattern x that we are classifying into one of the classes $\{1, \ldots, k\}$

$$\implies R_1 = \sum_{i=1}^{k} \mathbb{1}_{i\neq H_1(x)} p(c = i/x)p(x)$$

$$\implies R_1 = \sum_{i=1}^{k} (1 - \mathbb{1}_{i=H_1(x)}) p(c = i/x)p(x)$$

$$\implies R_1 = 1 - \sum_{i=1}^{k} \mathbb{1}_{i=H_1(x)} p(c = i/x)p(x)$$

We know $H_1(x) = c^*$ always. So we get,

$$R_1 = 1 - p(c = c^*/x)p(x)$$

where $c^* = \arg\max_c\ p(c/x)$.
Hence we get,

$$R_1 = 1 - max(p(c/x))p(x)$$

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:
(ii)

$$H_2(x) = c \sim p(c/x)$$
$$R_2 = E_{x,c,H_2(x)}[L] = E_x[E_{c,H_2(x)/x}[L]]$$

where $E_{c,H_2(x)/x}[L]$ denotes the expected loss over the joint distribution $(c, H_2(x))$ conditioned over x

$$\implies R_2 = E_x\left[\sum_{j=1}^{k}\sum_{i=1}^{k} L_{i,H_2(x)=j} p(c=j/x)p(c=i/x)\right]$$

$$\implies R_2 = E_x\left[\sum_{j=1}^{k}\sum_{i=1}^{k} \mathbb{1}_{i\neq j} p(c=j/x)p(c=i/x)\right]$$

Since there's only a single pattern x that we are classifying into one of the classes $\{1,\ldots,k\}$

$$\implies R_2 = \sum_{j=1}^{k}\sum_{i=1}^{k} \mathbb{1}_{i\neq j} p(c=j/x)p(c=i/x)p^2(x)$$

$$\implies R_2 = \sum_{j=1}^{k}\sum_{i=1}^{k} (1-\mathbb{1}_{i=j}) p(c=j/x)p(c=i/x)p^2(x)$$

$$\implies R_2 = 1 - \sum_{j=1}^{k}\sum_{i=1}^{k} \mathbb{1}_{i=j} p(c=j/x)p(c=i/x)p^2(x)$$

$$\implies R_2 = 1 - \sum_{i=1}^{k} p^2(c=i/x)p^2(x)$$

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:

(iii) Test 1 always perform better for the zero-one loss function because it chooses the class based on the MAP inorder which minimizes the Risk function R_1 better than Test 2 which chooses the class based on a random decision.
Whereas, both the tests behave the same when the posterior probability distribution is uniform (i.e.) $p(c = i/x)$ is equal $\forall i \in \{1,\ldots,k\}$. In this case both the risks R_1 and R_2 give the same result.

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:

Representing the equation in matrix format to simplify the expression. Let X represent the design matrix, Q represent the co-efficient matrix introducing the weights, w represent the parameter vector for the linear regression problem and let t represent the actual output.

$$X = \begin{bmatrix} \dots x_1^T \dots \\ \dots x_2^T \dots \\ \vdots \\ \dots x_n^T \dots \end{bmatrix}, t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}, Q = \begin{bmatrix} q_1 & 0 & \dots & 0 \\ 0 & q_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & q_n \end{bmatrix}, w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

Now we can represent the weighted error function $E_q(w)$ in matrix form as,

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

This is because,

$$(Xw - t)^T Q = \begin{bmatrix} q_1(x_1^T w - t_1) & q_2(x_2^T w - t_2) & \dots & q_n(x_n^T w - t_n) \end{bmatrix}$$

Also,

$$(Xw - t)^T Q = \begin{bmatrix} q_1(w^T x_1 - t_1) & q_2(w^T x_2 - t_2) & \dots & q_n(w^T x_n - t_n) \end{bmatrix}$$

$$(Xw - t)^T Q(Xw - t) = \begin{bmatrix} q_1(w^T x_1 - t_1) & q_2(w^T x_2 - t_2) & \dots & q_n(w^T x_n - t_n) \end{bmatrix} \begin{bmatrix} w^T x_1 - t_1 \\ w^T x_2 - t_2 \\ \vdots \\ w^T x_n - t_n \end{bmatrix}$$

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

Optimization :

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

$$E_q(w) = \frac{1}{2}w^T(X^T QX)w - (Y^T QX)w + \frac{1}{2}t^T Qt$$

Now we can optimize the above expression by finding $\nabla E_q(w)$

$$\nabla E_q(w) = \frac{1}{2}\left((X^T QX)^T + (X^T QX)\right)w - (t^T QX)^T$$

This can be simplified and rearranged because both Q and $(X^T QX)$ are symmetric matrices,

$$\nabla E_q(w) = (X^T QX)w - X^T Qt$$

The minimizer w^* of $E_q(w)$ satsifies $\nabla E_q(w) = 0$ and thus yields us the minima.

$$w^* = \arg\min_w \; E_q(w) = (X^T QX)^{-1}X^T Qt$$

(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:

I'm taking insights from part (a) and then representing the equation in matrix format to simplify the expression. Let Φ represent the design matrix, w represent the parameter vector for the ridge regression problem and let t represent the actual output.

$$\Phi = \begin{bmatrix} \dots \phi(x_1)^T \dots \\ \dots \phi(x_2)^T \dots \\ \vdots \\ \dots \phi(x_n)^T \dots \end{bmatrix}, t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}, w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

If the Hessian of a function is Positive Semi-definite, then the function is generally convex. Strictly convex nature is ensured when the Hessian is Positive definite. In matrix form we

can write the error function for ridge regression as,

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

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

Optimization :

To minimize the error function above we have to calculate the gradient of the error function (w.r.t) w which is $\nabla\widetilde{E}(w)$

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

We know that $(\Phi^T\Phi + \lambda I)$ is symmetric because both $\Phi^T\Phi$ and λI are symmetric. Hence, we can re-write $\nabla\widetilde{E}(w)$ as,

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

We can then also find the Hessian matrix as,

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

Checking for positive definiteness of $H(\widetilde{E}(w))$. Let us consider any $a \in \mathbb{R}^N$

$$a^T(\Phi^T\Phi + \lambda I)a = a^T\Phi^T\Phi a + \lambda a^T a = (\Phi a)^T(\Phi a) + \lambda(a^T)(a) \geqslant 0 = \|\Phi a\| + \lambda\|a\|$$

The above expression is always $\geqslant 0$ and the equality holds for the trivial case where $a = 0$. Hence, the above expression is > 0 for any general a other than the trivial case because it is the sum of 2 euclidean norms, which are always positive and hence the expression is always > 0 (provided $\lambda > 0$. So, the hessian of $\widetilde{E}(w)$ which is $H(\widetilde{E}(w))$ is positive definite. Hence the error function $\widetilde{E}(w)$ is convex.
We can find the minimizer w^* of the error function $\widetilde{E}(w)$ by equating it's gradient to zero and finding the corresponding w that satisfies it.

$$\nabla\widetilde{E}(w^*) = 0$$

$$\implies (\Phi^T\Phi + \lambda I)w^* - \Phi^T t = 0$$

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

Invertibility :

We got $(\lambda I + \Phi^T\Phi)$ as hessian of the error function $\widetilde{E}(w)$ and proved that it is a positive definite matrix. From Linear Algebra, we can say that the matrix $(\lambda I + \Phi^T\Phi)$ has all of it eigenvalues > 0 (also $\lambda > 0$) because it's positive definite .We also know from Linear Algebra that such a matrix is always an invertible matrix because it has all it eigenvalues > 0 and no eigenvalues as 0.
Hence, we proved and can state that the matrix $(\lambda I + \Phi^T\Phi)$ is invertible for $\lambda > 0$

(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:

We know that the minimizer of $E(w) = \frac{1}{2}||Xw - t||^2$ satisfies the equation $(X^TX)w = X^Tt$ We can only have a unique closed form solution for $w^* = (X^TX)^{-1}X^Tt$ when X and X^TX both are invertible matrices. In this question, we have

$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix}, X^TX = \begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix}, X^Tt = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

From the matrices we get $det(X) = det(X^TX) = 0$. This means that both X and X^TX are non-invertible and the solution to w^* is not unique and cannot be expressed as $w^* = (X^TX)^{-1}X^Tt$. Instead we have to solve $(X^TX)w = X^Tt$ and solve for possible solutions of w.

Let us consider w to be,

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

The equation $(X^TX)w = X^Tt$ can be expressed in matrix form as,

$$\begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

It can be simplified as,

$$\begin{bmatrix} 1 & -3 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

As a result we only get a single linear equation which represents a line in $w1, w2 \in \mathbb{R}^2$ plane,

$$w_1 - 3w_2 = -1 \tag{1}$$

Objective : To minimize the norm of $w = [w_1, w_2]^T$ under the condition (1). This is same as minimizing square of norm of w over the same condition.

$$f(w_1, w_2) = ||w||^2 = w_1^2 + w_2^2$$

Now, $f(w_1, w_2)$ can be written as a single variable function because $w_1 = 3w_2 - 1$. This gives us,

$$f(w_2) = (3w_2 - 1)^2 + w_2^2 = 10w_2^2 - 6w_2 + 1$$

Gradient of $f(w_2)$ (w.r.t) $= 0$ minimizes the function $f(w_2)$

$$\nabla f(w_2) = 20w_2 - 6 = 0$$

This gives us $w_2^* = \frac{3}{10}$ and $w_1^* = \frac{-1}{10}$ and thus,

$$w^* = \begin{bmatrix} w_1^* \\ w_2^* \end{bmatrix} = \begin{bmatrix} -0.1 \\ 0.3 \end{bmatrix}$$

Minimizing Regularization loss :

We know that the minimizer of $\widetilde{E}(w) = \frac{1}{2}||Xw - t||^2 + \frac{1}{2}||w||^2$ satisfies the equation $(X^TX + \lambda I)w = X^Tt$ We can only have a unique closed form solution for $w^* = (X^TX + \lambda I)^{-1}X^Tt$ when $(X^TX + \lambda I)$ is an invertible matrices. In this question, we have $\lambda = 1$ and

$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix}, (X^TX + I) = \begin{bmatrix} 6 & -15 \\ -15 & 46 \end{bmatrix}, X^Tt = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

The matrix $(X^TX + \lambda I)$ is invertible with positive eigenvalues as $1, 51$. This means that the solution to $w^* = (X^TX + \lambda I)^{-1}X^Tt$ is unique.

The equation $(X^TX + I)w = X^Tt$ can be expressed in matrix form as,

$$\begin{bmatrix} 6 & -15 \\ -15 & 46 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

Thus we get $w = [w_1, w_2]$ as,

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \frac{1}{51} \begin{bmatrix} 46 & 15 \\ 15 & 6 \end{bmatrix} \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

It can be simplified as,

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \frac{1}{51} \begin{bmatrix} -5 \\ 15 \end{bmatrix} = \begin{bmatrix} -5/51 \\ 15/51 \end{bmatrix}$$

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:

(a) Given K_1, K_2 are valid kernels functions mapping vectors from $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$$K_1(x, y) = \phi_1(x)^T \phi_1(y) \text{ and } K_1(x, y) = \phi_2(x)^T \phi_2(y)$$

where ϕ_1, $\phi_2 \in \mathbb{R}^d$ Then we have,

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

$$\implies K_3(x, y) = \phi_1(x)^T \phi_1(y) + \phi_2(x)^T \phi_2(y) + 7.5$$

If we can express $K_3(x, y)$ as the inner product of 2 vectors $\phi_3(x) \text{and} \phi_3(y)$ then it is a valid kernel. Let us consider a vector $\phi_3(x) = [\phi_1(x)^T \ \ \phi_2(x)^T \ \ \sqrt{7.5}]^T \in \mathbb{R}^{2d+1}$

$$\implies \phi_1(x)^T \phi_1(y) + \phi_2(x)^T \phi_2(y) + 7.5 = \begin{bmatrix} \phi_1(x)^T & \phi_1(x)^T & \sqrt{7.5} \end{bmatrix} \begin{bmatrix} \phi_1(y) \\ \phi_2(y) \\ \sqrt{7.5} \end{bmatrix}$$

$$\implies K_3(x, y) = \phi_1(x)^T \phi_1(y) + \phi_2(x)^T \phi_2(y) + 7.5 = \phi_3(x)^T \phi_3(y)$$

$\therefore K_3$ is a valid kernel function because it can be represented as an inner product $K_3(x, y) = \phi_3(x)^T \phi_3(y)$.

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

Solution:

(b) Let $K_1(x, y) = \sum_{i=1}^{d_1} \phi_{1i}(x)\phi_{1i}(y)$ and
$K_2(x, y) = \sum_{j=1}^{d_2} \phi_{2i}(x)\phi_{2i}(y)$

$$K_1(x, y)K_2(x, y) = \left[\sum_{i=1}^{d_1} \phi_{1i}(x)\phi_{1i}(y)\right] \left[\sum_{j=1}^{d_2} \phi_{2i}(x)\phi_{2i}(y)\right]$$

$$= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \phi_{4ij}(x)\phi_{4ij}(y)$$

where $\phi_4 : \mathbb{R}^d \to \mathbb{R}^{d_1 \times d_2}$ and $(\phi_{4ij} = \phi_{1i}\phi_{2j})\ \forall i \in \{1, \ldots, d_1\}$ and $\forall j \in \{1, \ldots, d_2\}$
Hence, K_4 such that $K_4(x, y) = K_1(x, y)K_2(x, y)$, is a valid kernel function representing the inner-product using a valid basis vector mapping ϕ_4.

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

Solution:

(c) Given $x, y \in \mathbb{R}^d$. Then, $x^T y = \langle x, y \rangle$ is a valid Kernel function because it represents the inner product of 2 vectors $\in \mathbb{R}^d$.

Let's consider $\phi_5 \in \mathbb{R}^{d+1}$

$$\phi_5(x) = \begin{bmatrix} x \\ 1 \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$\implies \phi_5(x)^T \phi_5(y) = \begin{bmatrix} x^T & 1 \end{bmatrix} \begin{bmatrix} y \\ 1 \end{bmatrix} = x^T y + 1$$

Hence, $(x^T y + 1)$ is a valid kernel function.

In part (b) we proved that product of any two valid kernels is also a valid kernel. In this question we have $K_5(x, y) = (x^T y + 1)^{73}$, which is the product of 73 valid kernels, and can further be decomposed into a single valid kernel function by multiplying 2 valid kernels each time to give rise to a valid kernel, while taking the product.

Thus we have proved that K_5 is a valid kernel function.

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

Solution:

(d) Given $K_1(x, y)$ and $K_2(x, y)$ to be valid kernel functions. So, we can say that K_1 and K_2 are both Positive Semi-Definite and symmetric.

Let us assume $K_2(x, y) = 3K_1(x, y)$

$$\implies K_6(x, y) = 6K_1(x, y) - 9K_1(x, y) = -3K_1(x, y)$$

We know that $K_1(x, y)$ is a valid kernel and K_1 is Positive Semi-Definite. But, $K_6(x, y) =$

$-3K_1(x, y)$ which means that the kernel matrix K_6 is Negative Semi-Definite.

Hence, K_6 is not a valid kernel function.

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

Solution:
(e)

$$2x^T y = (\sqrt{2}x)^T(\sqrt{2}y) = \langle \sqrt{2}x, \sqrt{2}y \rangle$$

where $x, y \in \mathbb{R}^d$

Hence, $2x^T y$ is a valid kernel function. We know that $exp(x) = \left[1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots\ldots\ldots \infty\right]$
Also, we proved that sum of valid kernel functions to be valid, in part(a), and product of valid kernel functions also to be valid, in part (b).

$$exp(2x^T y) = \left[1 + \frac{2x^T y}{1!} + \frac{(2x^T y)^2}{2!} + \frac{(2x^T y)^3}{3!} + \ldots\ldots\ldots \infty\right]$$

$\therefore$ We get $exp(2x^T y)$ to be the sum of a valid kernel functions. This is because we have $K(x, y) = \frac{(2x^T y)^n}{n!}$ to be product of valid kernel functions and hence valid.

Thus, K represented by $K(x, y) = exp(2x^T y)$ is a valid kernel function.

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

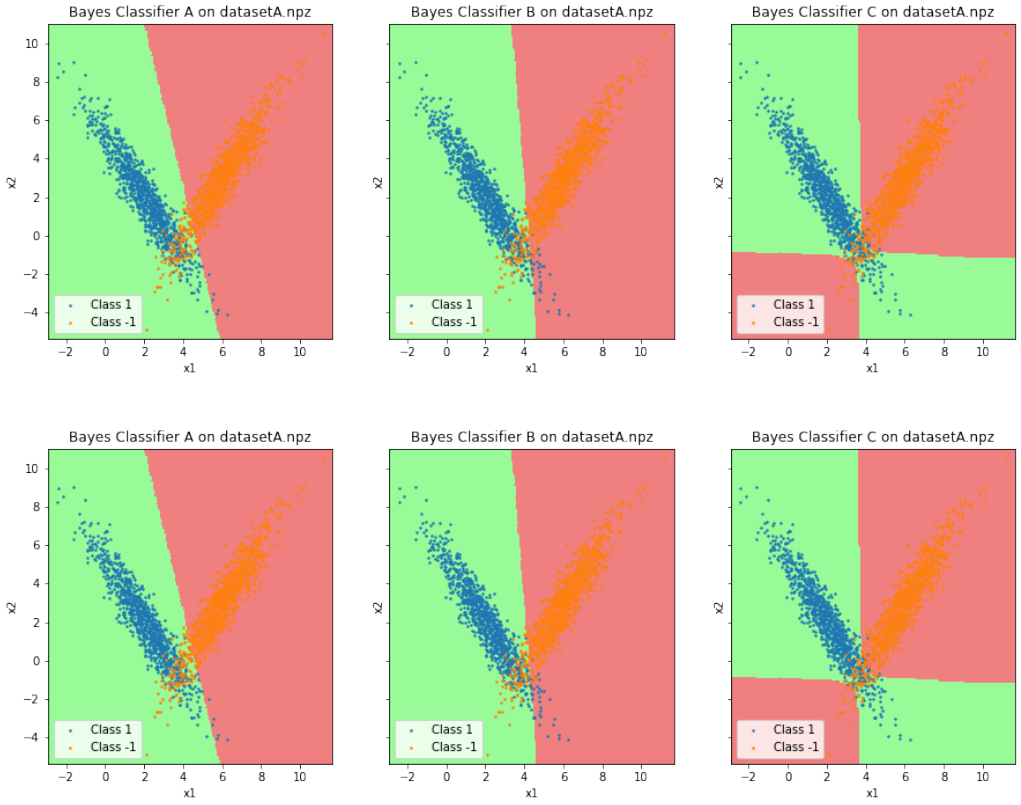
   Please see this folder for the template .ipynb file containing the helper functions, and you've to add the missing code to this file (specifically, three functions function_for_A, function_for_B, function_for_C and associated plotting/ROC code snippets) to implement the above three algorithms for the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

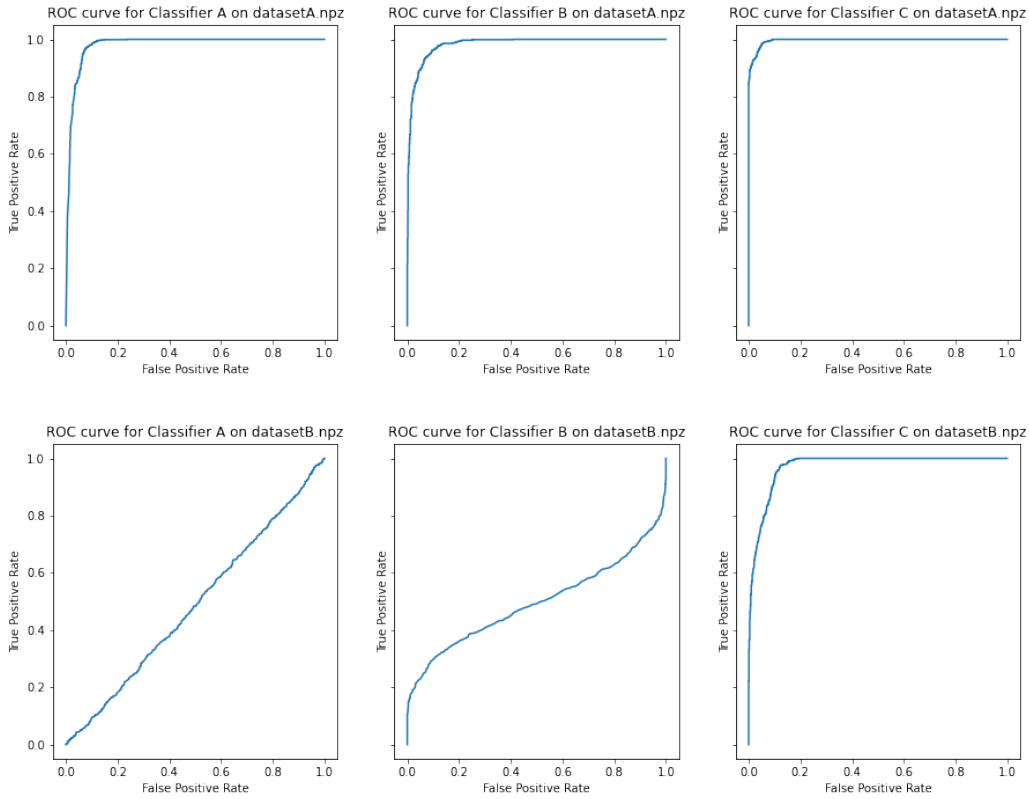
(a) (3 points) Plot all the classifiers (3 classification algorithms on 2 datasets = 6 plots) on a 2D plot, Add the training data points also on the plots. (Color the positively classified area light green, and negatively classified area light red as in Fig 4.5 in Bishop's book).

Solution:



(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:



(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	0.0335	0.0745

(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.

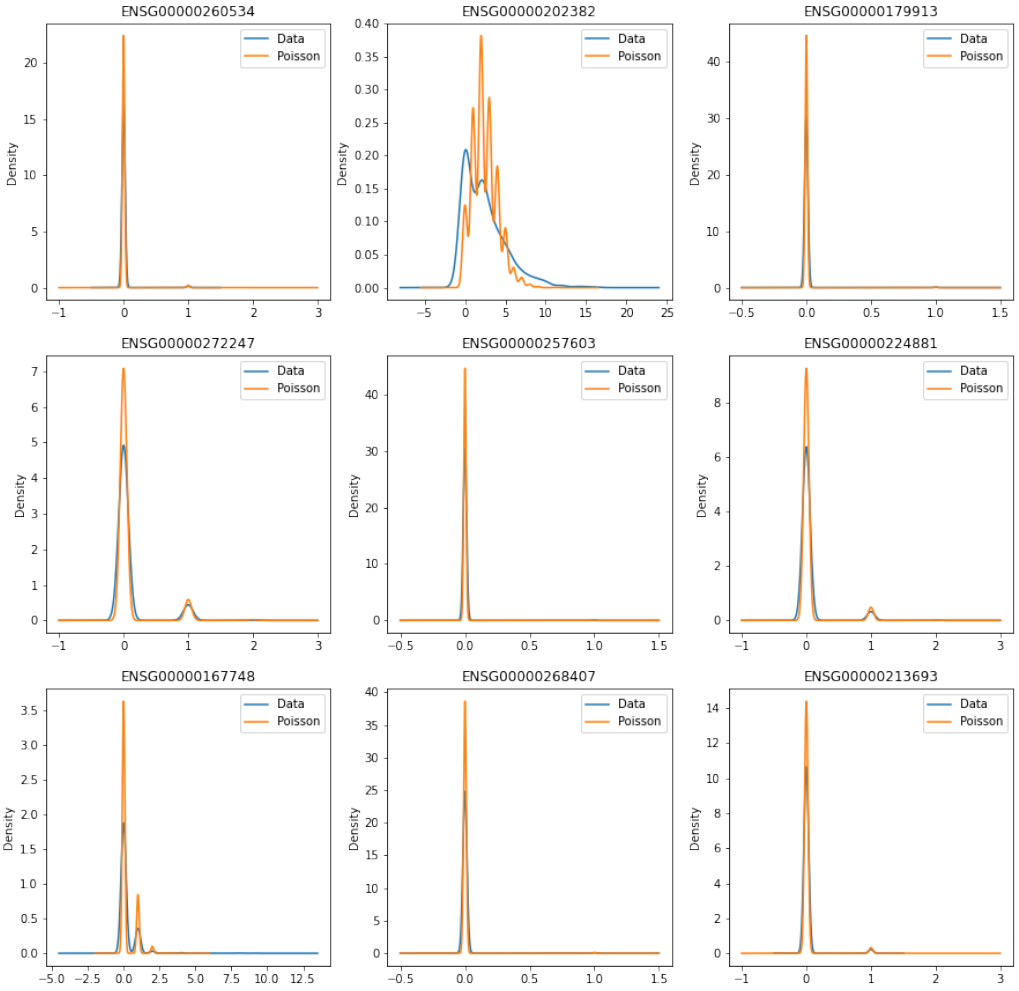
Solution: All 3 bayes classifiers perform reasonably well in dataset A. However, in dataset B, Bayes A and B have an error rate of around 50%, this implies that the covariance is different for both the classes therefore a more robust way was required to classify. Bayes C, which considered different variance performs well in both datasets.

In non parametric methods, no assumption about prior is required to be made. On the downside, choice of hyper-parameters is crucial and the model is very sensitive to it. Even though non parametric methods are computationally heavy, they perform very well in case of less datapoints, likewise in the given datasets.

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: The distribution is assumed to be poisson from visual inspection. It is verified by overlay of poisson distribution of same mean on various random gene's density distribution.



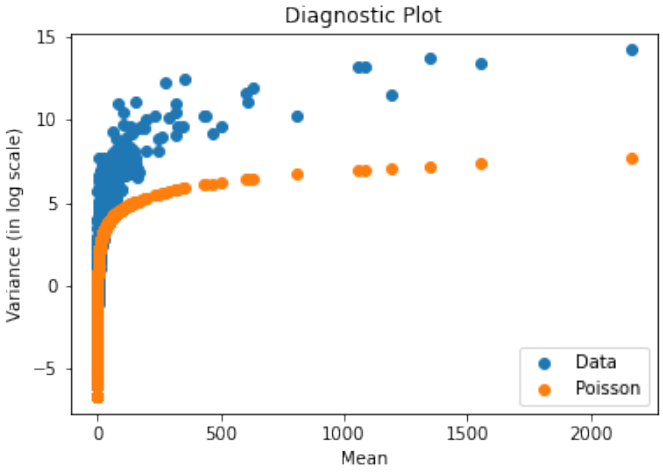
(b) (2 points) (MLE Code) How will you obtain the MLE estimates of the assumed model's parameters? (no need to derive it, just state your answer as a closed-form formula or as an optimization method). Write a code to estimate these parameters for each gene.

Solution: The assumed model is poisson distribution. The MLE estimator of λ is sample mean. The code for the same is submitted.

(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:



The variance of data is higher than the fitted poisson distribution, which hints a distribution with shape parameter could be a better fit. Conjugate priors to poisson distribution like gamma distribution could be tried as there is more flexibility for parameters and can fit better.