Solutions : **Name : Achyut Kulkarni**

1) **Density Estimation :**

(a)

Solution :

(i) Maximum Likelihood Estimator for Beta Distribution where $\beta = 1$ is :

We know that the PDF for Beta Distribution :

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad \text{where } B(\alpha,\beta) = \frac{\Gamma\alpha\,\Gamma\beta}{\Gamma(\alpha+\beta)} \;;\; \alpha > 0, \beta > 0$$

For $\beta = 1$ => we have $B(\alpha,\beta) = \frac{\Gamma\alpha}{\Gamma\alpha+1} = 1/\alpha$;

Now maximum likelihood estimation of the beta distribution after substitution is :

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \Rightarrow \alpha x^{\alpha-1}$$

Hence, the likelihood function is :

$$L(\alpha) = \sum_i^n ln(\alpha x^{\alpha-1})$$

To calculate the parameter α ; we have

$$\frac{\partial L(\alpha)}{\partial \alpha} = 0 \Rightarrow \frac{\partial}{\partial \alpha}\left(\sum_i^n ln(\alpha x^{\alpha-1})\right) = 0$$

$$\Rightarrow \frac{\partial}{\partial \alpha}\left(\sum_i^n ln\alpha + \sum_i^n ln(x)^{\alpha-1}\right) = 0$$

$$\Rightarrow \sum_i^n \frac{1}{\alpha} + (\alpha - 1)\sum_i^n ln(x) = 0$$

$$\Rightarrow \frac{N}{\alpha} + \sum_i^n ln(x) = 0$$

$$\Rightarrow \alpha = -\frac{N}{\sum_i^n ln(x)}$$

(ii) Maximum Likelihood Estimator for Normal Distribution :

We know that the PDF for Normal distribution is :

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ where } \mu = Mean\ and\ \sigma^2 = Variance$$

Now Mean and Variance = θ, for maximum likelihood :

Hence ;

Likelihood function for this is :

$$L(\theta) = \sum_i^n ln(N(\theta, \theta)) = \sum_i^n ln(\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\theta)^2}{2\theta}})$$

And to estimate the value of θ ;

We do $\frac{\partial}{\partial\theta}(L(\theta)) = 0$;

$$=> \frac{\partial}{\partial\theta}(\sum_i^n ln(\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\theta)^2}{2\theta}})) = 0 ;$$

$$=> \frac{\partial}{\partial\theta}(\sum_i^n (ln(\frac{1}{\sqrt{2\pi\theta}}) + ln(e^{-\frac{(x-\theta)^2}{2\theta}})) = 0 ;$$

$$=> \frac{\partial}{\partial\theta} (\frac{-1}{2} [\sum_i^n (ln(2\pi\theta) + \frac{(x_i-\theta)^2}{\theta}) = 0)$$

$$=> \frac{\partial}{\partial\theta}(\sum_i^n ln(2\pi) + \sum_i^n ln\theta + \sum_i^n \frac{(x_i-\theta)^2}{\theta}) = 0$$

$$=> (0 + \sum_i^n [\frac{1}{\theta} + \frac{-(x_i-\theta)^2}{\theta^2} - \frac{2(x_i-\theta)}{\theta}]) = 0$$

$$=> \frac{N}{\theta} + \frac{1}{\theta^2} \sum_i^n [-(x_i-\theta)^2 - 2\theta(x_i-\theta)] = 0$$

Expanding the terms and cancelling $2x_i\theta\ and\ \theta^2$ we get :

$$=> \frac{N}{\theta} + \frac{1}{\theta^2} \sum_i^n (\theta^2 - x_i^2) = 0$$

$$=> \frac{N}{\theta} + \frac{\theta^2}{\theta^2} \sum_i^n 1 - \frac{1}{\theta^2} \sum_i^n x_i^2 = 0$$

$$=> \frac{N}{\theta^2} + N - \frac{1}{\theta^2} \sum_i^n x_i^2 = 0$$

Multiplying θ^2 to both sides we get :

$$=> N\theta^2 + N\theta - \sum_i^n x_i^2 = 0$$

Hence by quadratic equation formula we get :

$$\theta = \frac{-N \pm \sqrt{N^2 + 4N(\sum_i^n x_i^2)}}{2N}$$

1)
(b)

Bias Estimation : Kernel Density Estimation

Let's Denote the KDE function as $E[f'(x)]$

We have $E[f'(x)] = \frac{1}{N}\sum_i^N E[\frac{1}{h}(\frac{x-Xi}{h})]$

$$=> \frac{1}{N}N\int \frac{1}{h}K(\frac{x-z}{h})f(z)dz$$

$$=> \int K(u)f(x-uh)du$$

Taylor Expansion :

$$f(x-uh) = f(x) + f^1(x)(-uh) + 1/2 * f^2(x)(-uh)^2 + ... + \frac{1}{n!}f^n(x)(-uh)^n + O(h^{n+1})$$

$$\int K(u)f(x-uh)du = f(x) + f^1(x)(-h)\int K(u)udu + ...$$

$$+ 1/2 * f^2(x)(-h)^2 \int K(u)u^2du + O(h^3)$$

$$= f(x) + 1/2 * f^2(x)h^2 \int K(u)u^2du + O(h^3)$$

Bias :

$$\text{Bias}(f'(x)) = E[f'(x)] - f(x) = 1/2 * f^2(x)h^2 \int K(u)u^2du + O(h^3)$$

2) Naive Bayes :
(a)
Logistic Regression assumes a parametric form for the distribution P(Y|X), then directly estimates its parameters from the training data. The parametric model assumed by Logistic Regression in the case where Y is boolean is:

$$P(Y = 1|x) = \frac{1}{(1+exp(w_0 + w_j X_j))}$$

$$P(Y = 0|x) = \frac{exp(w_0 + w_j X_j)}{1+exp(w_0 + w_j X_i)}$$

Y is boolean, governed by a Bernoulli distribution, with parameter π = P(Y = 1)

- X = (X_j.....X_n), where each X_j is a continuous random variable
- For each X_j , P(X_i |Y =y_k) is a Gaussian distribution of the form N(μ_{jk}, σ_j)

We now derive the parametric form of P(Y|X) that follows from this set of GNB assumptions. In general, Bayes rule allows us to write

$$P(Y = 1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)\ P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

Dividing both the numerator and denominator by the numerator yields:

$$P(Y = 1|X) = \frac{1}{(1+ (P(Y{=}0)P(X|Y{=}0))/(P(Y{=}1)P(X|Y{=}1)))}$$

or equivalently

$$P(Y = 1|X) = \frac{1}{(1+exp(ln((P(Y{=}0)P(X|Y{=}0))/(P(Y{=}1)P(X|Y{=}1)))))}$$

Because of our conditional independence assumption we can write this

$$P(Y = 1|X) = \frac{1}{(1+exp(ln((P(Y{=}0))/(P(Y{=}1)\ )) + \sum_j ln((P(X_j\,|Y{=}0))/(P(X_j\,|Y{=}1)))))} \quad \text{........ (1)}$$

$$\Rightarrow P(Y = 1|X) = \frac{1}{(1+exp(ln((1-\pi)/\pi) + \sum_j ln((P(X_j\,|Y{=}0))/(P(X_j\,|Y{=}1)))))} \quad \text{...... (2)}$$

Note the final step expresses P(Y = 0) and P(Y = 1) in terms of the binomial parameter π

Now consider just the summation in the denominator of equation (1). Given our assumption that P(X_j |Y = Y_k) is Gaussian, we can expand this term as follows:

$$\sum_j ln(\frac{(P(X_j|Y=0)}{P(X_j|Y=1}) = \sum_j ln(\frac{(1/\sqrt{2\pi\sigma_j^2})exp(-(X_j-\mu_{i0})^2/(2\sigma_j^2)))}{(1/\sqrt{2\pi\sigma_j^2})exp(-(X_j-\mu_{i1})^2/(2\sigma_j^2)))}) \quad \text{..... (3)}$$

$$\sum_j lnexp\ (\frac{((X_j-\mu_{j1})^2-(X_j-\mu_{j0})^2}{(2\sigma_j^2)})$$

Expanding the above equation ..using $(a+b)^2 = a^2 + b^2 + 2ab$ and simplifying….. We get

$$=> \sum_j \frac{2X_j(\mu_{j0} - \mu_{j1}) + \mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}$$

$$=> \sum_i ((\frac{(\mu_{i0} - \mu_{i1})X_j}{\sigma_j^2}) + \frac{\mu_{j1}^2 - \mu_{j0}^2}{2\sigma_j^2})$$

Substituting this in (2) we get

$$P(Y = 1|X) = \frac{1}{(1+exp(ln((1-\pi)/\pi) + \sum_j(((\mu_{j0} - \mu_{j1})X_j/\sigma_j^2) + (\mu_{j1}^2 - \mu_{j0}^2/(2\sigma_j^2)))}$$

Or equivalently

$$P(Y = 1|X) = \frac{1}{(1+exp(w_0 + \sum_j^n w_j X_j))}$$

where the weights w1 ...wn are given by :

$$w_j = \frac{\mu_{j0} - \mu_{j1}}{\sigma_j^2}$$

$$w_0 = ln(\frac{1-\pi}{\pi}) + \sum_j \frac{(\mu_{j1}^2 - \mu_{j0}^2)}{2\sigma_j^2}$$

Are the parameters of the above equation

(2) Naive Bayes :

(b)

Solution:

To estimate $\mu_{jk}\, and\, \sigma_j^2$ **we write the likelihood function as :**

$$L = log \prod_{k=0}^{1} \prod_{i=1, y_i = k}^{N} \prod_{j=1}^{D} \left[(2\pi\sigma_j^2)^{-1/2} exp \left\{ \frac{-1}{2\sigma_j^2}(x_{ji} - \mu_{jk})^2 \right\} \right]$$

$$=> -\sum_{k=0}^{1} \sum_{i=1, y_i = k}^{N} \sum_{j=1}^{D} \left[\frac{1}{2} log(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2}(x_{ji} - \mu_{jk})^2 \right]$$

For μ_{jc} :

$$\frac{\partial L}{\partial \mu_{jk}} = \sum_{i=1, y_i=k}^{N} \frac{x_{ji} - \mu_{jk}}{\sigma_j 2} = 0$$

$$\sum_{i=1, y_i=k}^{N} x_{ji} = \sum_{i=1, y_i i=k}^{N} \mu_{jk} = N_k \mu_{jk}$$

$$\mu_{jk} = \frac{1}{N_k} \sum_{i=1, y_i=k}^{N} x_{ji}$$

Now for σ_j :

$$\frac{\partial L}{\partial \sigma_j^2} = \sum_{k=0}^{1} \left[- \sum_{i=1, y_i=k}^{N} \frac{1}{2\sigma_j^2} + \sum_{i=1, y_i=k}^{N} \frac{1}{2\sigma_j^4} (x_{ji} - \mu_{jk})^2 \right] = 0$$

$$\sum_{k=0}^{1} \sum_{i=1, y_i=k}^{N} \frac{1}{\sigma_j^2} = \sum_{k=0}^{1} \sum_{i=1, y_i=k}^{N} \frac{1}{\sigma_j^4} (x_{ji} - \mu_{jk})^2$$

$$\sigma_j^2 = \frac{1}{N} \sum_{k=0}^{1} \sum_{i=1, y_i=k}^{N} (x_{ji} - \mu_{jk})^2$$

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{k=0}^{1} \sum_{i=1, y_i=k}^{N} (x_{ji} - \mu_{jk})^2}$$

Now for π :

W.k.t : for Bernoulli Distribution :

Given : $\pi = P(Y = 1) \Rightarrow P(Y = 0) = (1 - \pi)$

The likelihood function is :

$$L = \sum_{i}^{n} \pi^{y_i} (1 - \pi)^{(1 - y_i)}$$

The log likelihood function is :

$$LL = log \sum_{i}^{n} \pi^{y_i} (1 - \pi)^{(1 - y_i)} ;$$

Has to equated to zero to get π *value*

$$\frac{\partial LL}{\partial \pi} = \frac{\partial}{\partial \pi}\left(\sum_i^n y_i log\pi + (1-y_i)log(1-\pi)\right) = 0$$

$$=> \sum_i^n yi * \frac{1}{\pi} - \sum_i^n \frac{(1-yi)}{1-\pi} = 0$$

$$=> \sum_i^n yi * \frac{1}{\pi} = \sum_i^n \frac{1}{1-\pi} - \sum_i^n \frac{y_i}{1-\pi}$$

$$=> \sum_i^n yi * \frac{1-\pi}{\pi} = N - \sum_i^n yi$$

$$=> \left(\frac{1-\pi}{\pi}\right) = \frac{N}{\sum_i^n yi} - \frac{\sum_i^n yi}{\sum_i^n yi}$$

$$=> \frac{1}{\pi} - 1 = \frac{N}{\sum_i^n yi} - 1$$

$$\pi = \frac{\sum_i^n yi}{N}$$

(3)

(a) KNN Theory :

Given : Point is (20,7).

We calculate Mean and Standard Deviation of the points with X and Y co-ordinates respectively.

$\mu_x = 12.769$; $\mu_y = 12.307$; $\sigma_x = 20.716$; $\sigma_y = 25.930$

X	**Y**	**Manhattan_distance**	**Euclidean_distance**
0	49	2.585099	1.885585
-7	32	2.267391	1.621126
-9	47	2.942397	2.083036
29	12	0.627249	0.475297
49	31	2.325366	1.678133
37	38	2.016081	1.450025
8	9	0.656365	0.584348

13	-1	0.646403	0.457548
-6	-3	1.640655	1.312925
-21	12	2.171877	1.988426
27	-32	1.8419	1.5415
19	-14	0.858123	0.81129
27	-20	1.379127	1.094691

(i) Manhattan Distance (L1):

- **K=1 :**

(29,12) has the shortest Manhattan Distance => (20,7) is closest to (29,12) which belongs to Electrical Engineering.
The class which (29,12) belongs to is thus **Electrical Engineering**.

- **K=5 :**

[(29,12),(13,-1),(8,9),(19,-14),(27,-20)] are the 5 closest points to (29,12). Which belong to [(EE),(CS),(CS),(Eco),(Eco)] respectively.There are 2 Majors (Economics and Computer Science) with 2 neighbours each as nearest. But since nearest amongst them is Computer Science (13, -1) we choose **Computer Science** as major.

(ii) Euclidean Distance(L2) :

- **K=1 :**

(13,-1) has the shortest euclidean distance => (20,7) and (13,-1) belongs to Computer Science. Hence, (20,7) belongs to class **Computer Science**

- **K=5 :**

[(13,-1)(29,12)(8,9),(19,-14),(27,-20)] are the closest points to (20,7) and they belong to [(CS),(EE),(CS),(Eco),(Eco)] respectively.There are 2 Majors (Economics and Computer Science) with 2 neighbours each as nearest. But since nearest amongst them is Computer Science (13, -1) we choose **Computer Science** as major.

<u>Comparison of results :</u>

From the above results we can see that for Euclidean Distance , the point is classified as a point belonging to Computer Science. The tie breaker between Economics and CS is broke by the closest distance of a point (13,-1) belonging to CS .

However in Manhattan Distance , the classification for K=1 and K=5 varies as at K=1 it belongs to Electrical Engineering whereas at K=5 there is a tie and is broken by a point closest to Computer Science(CS). Hence the point belongs to CS at K=5.

3) Nearest Neighbour Problem :

(b)

(i) Using the fact that P c Kc = K, derive the formula for unconditional density p (x).

Solution :

Given : p (x | Y = c) = $\frac{Kc}{N_c V}$;

p (Y = c) = $\frac{N_c}{N}$;

P(x) = ?

We know that :

$$P(x) = \sum_{i}^{c} P(x|y) * P(y)$$

$$=> \sum_{i}^{c} \frac{K_i}{N_i V} * \frac{N_i}{N}$$

$$=> \sum_{i}^{c} \frac{K_i}{NV}$$

But we know that $\sum_{i}^{c} K_i = K$. Therefore the above equation becomes

$$=> \frac{K}{NV}$$

$$=> P(x) = \frac{K}{NV}$$

(ii) Using Bayes rule, derive the formula for the posterior probability of class membership p (Y = c | x).

From Naive Bayes Theorem we have :

$$P(Y = c|X) = \frac{P(X|Y=c)*P(Y=c)}{P(X)}$$

$$P(Y = c|X) = \frac{\frac{Kc}{N_c V} * \frac{N_c}{N}}{\frac{K}{NV}}$$

$$P(Y = c|X) = \frac{K_c}{K}$$

4) Decision Trees :
(a)

Solution:

Here we are considering Log to base 2 : i.,e $loga = log_2 a$
$p = 23 + 50 = 73$ and $n = 5 + 22 = 27$

$$H(\frac{p}{p+n}, \frac{n}{p+n}) = -\frac{73}{100}log(\frac{73}{100}) - \frac{27}{100}log(\frac{27}{100}) = 0.84$$

$I(weather) = 0.84 - EH(weather)$
$= 0.84 - [28/100 * H(23/28 , 5/28) + 72/100 * H(50/72, 22/72)]$
$= 0.84 - [28/100 * H(23/28 , 5/28) + 72/100 * H(50/72, 22/72)]$

Where $H(23/28, 5/28) = -23/28 * log(23/28) - 5/28 * log(5/28)$
<=> $H(23/28, 5/28) = 23/28 * log(28/23) + 5/28 * log(28/5)$ > 0 … (1)
$(since.. log\frac{a}{b} > 0 \; if \; a > b)$
And $H(50/72, 22/72) = -50/72 * log(50/72) - 22/72 * log(22/72)$
<=> $H(50/72, 22/72) = 50/72 * log(72/50) + 22/72 * log(72/22)$ > 0 …(2)
$(since.. log\frac{a}{b} > 0 \; if \; a > b)$

$I(weather) = 0.84 - K$; where $K =$
$[28/100 * H(23/28 , 5/28) + 72/100 * H(50/72, 22/72)]$ and $K >$ 0 from (1) and (2)

Now,
$I(Traffic) = 0.84 - EH(Traffic)$
$= 0.84 - [73/100 * H(73/73 , 0) + 27/100 * H(0, 27/27)]$
$= 0.84 - 0$
$= 0.84$

Here, $0.84 > (0.84 - K)$ as $K > 0$
Hence , we can say
$I(Traffic) > I(weather)$
Therefore, we choose ***Traffic*** attribute as the first node to split

(b)
Solution :

There is no difference in the classification or any decision made by the trees. Normalisation is basically done to scale all parameters to same level in order to maintain fair comparison between them.
Also whenever the data is normalised the values are distributed in range (0,1) however the values are just scaled relatively and has no impact on the classification ability of decision making tree.

(c)

Solution :
Here we denote $logx$ as the log base 2 equivalent i.,e $logx \Leftrightarrow log_2 x$
We know that Gini Index :

$$\sum_{k=1}^{K} p_k(1-p_k) \text{ ---- (1)}$$

And cross entropy is :

$$-\sum_{k=1}^{K} p_k log p_k \text{ -----(2)}$$

We have to show 1<2 => (1) - (2) <0 i.,e

$$\sum_{k=1}^{K} p_k(1-p_k) + \sum_{k=1}^{K} p_k log p_k < 0$$

$$=> \sum_{k=1}^{K} p_k(1-p_k+logp_k) < 0$$

$$=> (1-p_k+logp_k) < 0 \textit{ needs to be satisfied}$$

$$=> (1-x+logx) \text{ --- (3) from above}$$

W.k.t for all K's $0 \le p_k \le 1$. Therefore we can find the critical point by differentiating the equation (3) w.r.t x i.,e

$$\frac{d}{dx}(1-x+logx) = -1 + \frac{1}{xlog(2)} \text{ which is continuous and >0}$$

The Function $\frac{1}{xlog(2)} \leq 1 \ \forall \ x \ \in (0, 1)$ and function (3) is $0 \ for \ x = 1 \ and \ x -> 0^{+} (3) \ -> \ -\infty$ and for any value between $x = 0 \ and \ x = 1, \ the \ function \ is \ less \ than \ zero.$ Also if we consider (3) to have a positive critical point , we must have $\frac{d}{dx}(1 - x + logx)$ <0 which is incorrect as it is > 0 . hence, .

$$(1 - x + logx) < 0$$

This proves that Gini Index < Cross Entropy .

(5)

(a) Data Inspection :

- There are 10 attributes . Column 1 to Column 10 are the attributes
- Here, all attributes are not used for classification. The first column is an Id number column. This data is only to keep the count and identify the data point. However this point does not provide us any information about the classification as it does not provide us any information for classifying the data points
- There are 6 classes
- Here the number of occurrences of each classes in the data set is not following any Uniform distribution curve. Ie., The values of number of each classes are random and do not follow any order and hence we can say that the distribution is not Uniform.The class with majority is 2. As the number of data points on class 2 is maximum

(5)

(**c) Performance Comparison :**

- **KNN :**

Testing Accuracy ----- >>>>

Manhattan Distance Accuracy (L1) ----->>>
accuracy for K = 1 = 61.1111111111 %
accuracy for K = 3 = 61.1111111111 %
accuracy for K = 5 = 55.5555555556 %
accuracy for K = 7 = 55.5555555556 %
Euclidean Distance Accuracy (L2)------>>>
accuracy for K = 1 = 61.1111111111 %
accuracy for K = 3 = 61.1111111111 %
accuracy for K = 5 = 55.5555555556 %
accuracy for K = 7 = 55.5555555556 %

Training Accuracy with LOO ---->>>

Manhattan Distance Accuracy (L1)----->>
accuracy for K = 1 = 74.4897959184 %
accuracy for K = 3 = 71.9387755102 %
accuracy for K = 5 = 71.4285714286 %
accuracy for K = 7 = 68.8775510204 %
Euclidean Distance Accuracy (L2)----->>
accuracy for K = 1 = 73.9795918367 %
accuracy for K = 3 = 68.3673469388 %
accuracy for K = 5 = 68.3673469388 %
accuracy for K = 7 = 66.8367346939 %

- **Naive Bayes :**

Testing Accuracy : 33.33 %

Training Accuracy : 54.59%

- **Discussion :**

From the above results we can confirm that the results from Knn and Naive bayes are not similar for the same data sets. We can see also see that Knn provides higher accuracy for training and testing data when compared with Naive Bayes Algorithm.
I think in the data set provided to us, for the Naive Bayes' Algorithm we calculate Prior Probabilities of each class and since the data set is small and the classes have significant examples, the value of the prior probabilities add more weight to the evaluation of Probability of a data point which creates an unnecessary bias towards or against a certain class and thus this causes hindrance in the accuracy of the Naive Bayes Algorithm. However, since KNN is a doesn't carry the prior probabilities it has better accuracy.

This also means that as Naive Bayes' is a generative classifier , it predicts the class based on some assumptions made from the data i.,e the by knowing the prior probabilities thus generalising how the data could have been produced rather than like a discriminative classifier (KNN) which does not bother about how the data is produced as it does not make any generalisations or assumptions prior to the availability of the data point and just discriminates as and when the data comes.

We also can check the performance of the Naive Bayes and KNN classifier using an inbuilt timer. From the observations made the Naive Bayes' Classifier is almost 28 times faster than KNN for the current data set for my program.
After multiple observations it was evident that the performance of Naive Bayes' was higher than KNN classifier.