USN | | | | | | | | | | |

Internal Assessment Test II –JULY 2024

Sub	Data Science and Its Applications					SubCode:	21AD62	Branch:	AIML/AInDS
Date	9/7/24	Duration:	90 minutes	Max Marks:	50	SEC	VI-A		OBE

Scheme and Solutions

		MARKS	CO	RBT

1a)	What is the need of logistic function in logistic regression and explain the maximum likelihood estimation and goodness fit of it.
	Answer: -
	Need for logistic function-2M
	`def logistic_prime(x):`
	`  return logistic(x) * (1 - logistic(x))`
	$y_i = f(x_i\beta) + \varepsilon_i$
	where *f* is the logistic function.
	We need the output value as 1 or zero.
	
	Maximum Likelihood estimation-4M
	Likelihood calculation, Negative likelihood estimation
	$\log L(\beta \mid x_i, y_i) = y_i \log f(x_i\beta) + (1 - y_i) \log (1 - f(x_i\beta))$
	Because log is strictly increasing function, any beta that maximizes the log likelihood also maximizes the likelihood, and vice versa.
	`def logistic_log_likelihood_i(x_i, y_i, beta):`
	`  if y_i == 1:`
	`  return math.log(logistic(dot(x_i, beta)))`
	`  else:`
	`  return math.log(1 - logistic(dot(x_i, beta)))`
	
	Goodness of fit-4M
	Python code to calculate precison, recall.
	`true_positives = false_positives = true_negatives = false_negatives = 0`
	`for x_i, y_i in zip(x_test, y_test):`
	`  predict = logistic(dot(beta_hat, x_i))`
	`  if y_i == 1 and predict >= 0.5:  # TP: paid and we predict paid`
	`  true_positives += 1`
	`  elif y_i == 1:  # FN: paid and we predict unpaid`
	`  false_negatives += 1`
	`  elif predict >= 0.5:  # FP: unpaid and we predict paid`
	`  false_positives += 1`
	`  else:  # TN: unpaid and we predict unpaid`
	`  true_negatives += 1`
	`precision = true_positives / (true_positives + false_positives)`
	`recall = true_positives / (true_positives + false_negatives)`
		10	3	L2

2a)	Explain SVM and the need of Kernel in SVM. **Answer: -** **SVM-2M** **SVM is a supervised learning algo. It is used for classification. To find the best line or decision boundary that can segregate n dimensional space.** **Need of Kernel-2M** **To map points into high dimensional space to find the hyperplane+ diagrams**	4	3	L2

2b)	Differentiate Multiple linear regression with simple linear regression. Write a python code to build both models and compute their R-squared value.
	Answer: -
	Difference between multiple and simple linear regression-2M
	In simple linear we will have one independent variable and multiple linear we will have multiple independent variables+ model equations.
	$y_i = \beta x_i + \alpha + \varepsilon_i$
	$y_i = \alpha + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i$
	Building both models-2M
	Python code to generate models
	`def predict(alpha, beta, x_i):`
	`  return beta * x_i + alpha`
	`def predict(x_i, beta):`
	`  """assumes that the first element of each x_i is 1"""`
	`  return dot(x_i, beta)`
	R-squared value-2M
	Python code to compute R-squared=1-sum of square errors/sum of mean errors.
	`def total_sum_of_squares(y):`
	`  """the total squared variation of y_i's from their mean"""`
	`  return sum(v ** 2 for v in de_mean(y))`
	`def r_squared(alpha, beta, x, y):`
	`  """the fraction of variation in y captured by the model, which equals`
	`  1 - the fraction of variation in y not captured by the model"""`
	`  return 1.0 - (sum_of_squared_errors(alpha, beta, x, y) /`
	`  total_sum_of_squares(y))`
		6	3	L2

3a)	List the steps of K-Nearest Neighbors algorithm and write a python code to classify the IRIS dataset using K-Nearest Neighbors. **Answer: -** **Steps of K-Nearest Neighbors-3M** • Step 1: Selecting the optimal value of K. K represents the number of nearest neighbors that needs to be considered while making prediction. • Step 2: Calculating distance. • Step 3: Finding Nearest Neighbors. • Step 4: Voting for Classification or Taking Average for Regression. **IRIS dataset python program-3M**	6	3	L3

3b) Explain the following i) Curse of Dimensionality ii) F1-Score iv) Lasso Regression v) Bias

Answer: -

Curse of Dimensionality-1M

The data set contains many attributes which may lead to overfitting and classifier performance is reduced.

F1-Score-1M

Its an evaluation metric.It is harmonic mean of precision and recall

$$F1\ Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Lasso Regression -1M

Lasso adds a penalty term as lambda*slope, L1 regularization, can make the coefficients as 0.

Bias-1M

A difference between the predicted and expected value.

4 | 3 | L2

4a) Apply K-means algorithm for K=2 where initial cluster centers are (1, 1) and (5, 7). Execute for two iterations.

Answer: -

First Iteration: - Cluster1 {R1, R2, R3} and Cluster2 {R4, R5, R6, R7} -----5M

Iteration1:

Record Number	Close to C1(1.0, 1.0)	Close to C2(5.0, 7.0)	Assign to cluster
R1(1.0,1.0)	dist(R1, C1)=0.0	dist(R1, C2)=7.21	Cluster1
R2(1.5,2.0)	dist(R2, C1)=1.12	dist(R2, C2)=6.12	Cluster1
R3(3.0,4.0)	dist(R3, C1)=3.61	dist(R3, C2)=3.61	Cluster1
R4(5.0,7.0)	dist(R4, C1)=7.21	dist(R4, C2)=0.0	Cluster2
R5(3.5,5.0)	dist(R5, C1)=4.12	dist(R5, C2)=2.5	Cluster2
R6(4.5,5.0)	dist(R6, C1)= 5.31	dist(R6, C2)=2.06	Cluster2
R7(3.5,4.5)	dist(R7,C1)=4.30	dist(R7, C2)=2.92	Cluster2

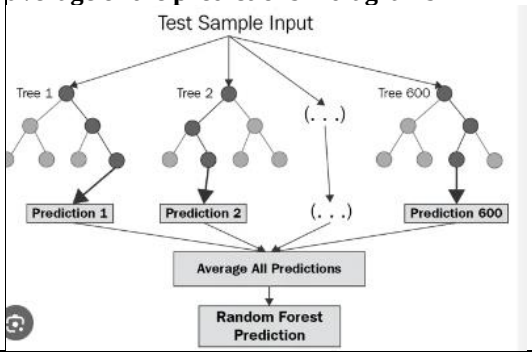
Second Iteration: -Cluster1 {R1, R2} and Cluster2 {R3, R4, R5, R6, R7}. ----------5M

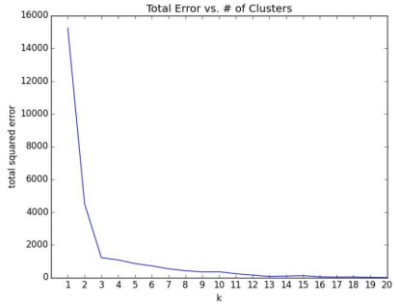
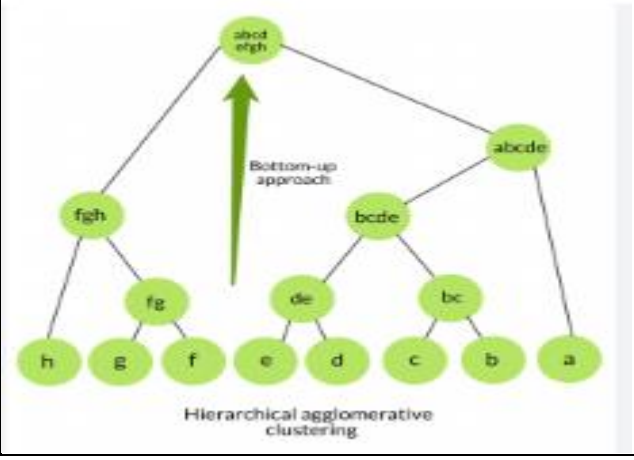
Iteration2:

Record Number	Close to C1(1.83, 2.33)	Close to C2(4.12, 5.37)	Assign to cluster
R1(1.0,1.0)	dist(R1, C1)=1.57	dist(R1, C2)=5.37	Cluster1
R2(1.5,2.0)	dist(R2, C1)=0.47	dist(R2, C2)=4.27	Cluster1
R3(3.0,4.0)	dist(R3, C1)=2.04	dist(R3, C2)=1.77	Cluster2
R4(5.0,7.0)	dist(R4, C1)=5.64	dist(R4, C2)=1.85	Cluster2
R5(3.5,5.0)	dist(R5, C1)=3.15	dist(R5, C2)=0.72	Cluster2
R6(4.5,5.0)	dist(R6, C1)=3.78	dist(R6, C2)=0.53	Cluster2
R7(3.5,4.5)	dist(R7,C1)=2.74	dist(R7, C2)=1.07	Cluster2

X1	1	1.5	3	5	3.5	4.5	3.5
X2	1	2	4	7	5	5	4.5

10 | 4 | L3

5a)	What is the need of calculating entropy and information gain in decision trees? Calculate the entropy value for P(yes) is 6/10 and P(no) is 4/10 given yes, no are class labels. **Answer: -** **Entropy-1.5M** **Its is the impurity in the dataset. General case of entropy** **Information gain 1.5M** **As entropy increases the information gain reduces. So the splitting attribute is selected based on the max info gain.** **Calculate Entropy-2M** (Entropy = -(4/10) * log2(4/10) - (6/10) * log2(6/10 (0.736966-) * 0.6 - (1.32193-) * 0.4- = 0.442179 + 0.52954 = 0.971719	5	4	L3
5b)	Explain how ensemble methods like bagging, random forest help in getting more accurate predictions **Answer:-** **Bagging 2.5M** **The training set is splitted and given to the same model and it takes the average of all.** **Random Forest-2M.** **The training dataset constructs multiple decision trees which are combined, and it takes the average of the predictions + diagrams** 	5	4	L2

6a)	How to choose K value in K-means and explain bottom-up hierarchical clustering approach with an example in detail. **Answer:-** **How to choose K value-2M** **Plotting the sum of squared errors (between each point and the mean of its cluster) as a function of k and looking at where the graph "bends".**  **Bottom-up hierarchical clustering-4M** Step 1: Compute the proximity matrix using a particular distance metric. Step 2: Each data point is assigned to a cluster. Step 3: Merge the clusters based on a metric for the similarity between clusters. Step 4: Update the distance matrix. **Example-4M**  	10	4	L2