


Final Assessment Test – November 2024

Course: BCSE334L - Predictive Analytics

Class NBR(s): 1761/1767/1771/1777/1780

Time: Three Hours

Slot: E1+TE1

Max. Marks: 100

- KEEPING MOBILE PHONE/ELECTRONIC DEVICES EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
 ➤ DON'T WRITE ANYTHING ON THE QUESTION PAPER

 Answer ALL Questions

(10 X 10 = 100 Marks)

- Compare and contrast predictive and descriptive models in terms of methodology, data requirements, and outcomes. Provide two examples of their applications in finance and healthcare.
- Given below the dataset, which contains test scores of five students across various subjects, use the K-Nearest Neighbors (KNN) method to impute the missing values with $k=3$. Apply Spatial Sign Transformation (SST) to handle potential outliers and find the top outlier.

Student	Math	Science	English	History	Art
1	85	78	---	92	88
2	72	----	85	80	75
3	90	88	92	85	85
4	65	70	75	78	---
5	----	85	80	89	82

- Differentiate between k-fold cross-validation and leave-one-out cross-validation. In what cases would leave-one-out cross-validation be more suitable than k-fold, and how do both methods influence bias and variance? Explain with examples appropriately.
- Given below the utility matrix, calculate the similarity between:
 - Users X and Y
 - Users X and Z

Use Jaccard Similarity, Cosine Similarity, and Pearson Correlation Coefficient to compute the similarities. Compare the results and explain which users have more similar ratings based on these similarity measures.

User	M1	M2	M3	M4	M5
X	5	4	3	---	2
Y	4	5	3	---	3
Z	---	2	---	5	3
W	3	3	---	4	---

- Using Support Vector Regression (SVR) with a linear kernel, predict the output y for the following dataset:

$$x_1 = 1, y_1 = 4$$

$$x_2 = 2, y_2 = 6$$

$$x_3 = 3, y_3 = 8$$
 Formulate the true relationship. Set up the SVR model and use it to predict y when $x=4$. Compare the SVR prediction with the true relationship, calculate residuals and errors for each data point, and discuss how the SVR model handles the deviations from the actual output.

6. Based on the given dataset, build a Regression Decision Tree to predict children's total playtime (in minutes) at a park. Predict the total playtime for a child with the following attributes:

Weather: Sunny, Age Group: Kid, Day Type: Weekend, Is School Holiday: No.

Child ID	Weather	Age Group	Day Type	Is School Holiday	Total Playtime(minutes)
1	Sunny	Toddler	Weekday	No	40
2	Cloudy	Kid	Weekend	Yes	70
3	Rainy	Toddler	Weekday	No	20
4	Sunny	Kid	Weekend	Yes	80
5	Cloudy	Teenager	Weekday	No	50
6	Sunny	Teenager	Weekend	Yes	90
7	Rainy	Kid	Weekday	No	30
8	Sunny	Toddler	Weekend	Yes	50
9	Cloudy	Kid	Weekday	No	60
10	Rainy	Teenager	Weekend	Yes	55

7. The vectors $(2, 0, 1, -1)$ and $(1, 1, 1, 1)$ belong to the class (target value = 1). The vectors $(0, 1, -1, -1)$ and $(-1, -1, -1, 1)$ do not belong to the class (target value = -1). Assume the learning rate is $\eta=1$, and the initial weights are $w=(0,0,0,0)$. Apply the perceptron learning rule to find the final weights after training on the given dataset.
8. Consider the following table with attributes shape, color, and edible (whether a fruit is edible or not):

S.No	Shape	Color	Edible
1	Round	Red	Yes
2	Round	Green	No
3	Oval	Red	Yes
4	Oval	Green	No
5	Round	Yellow	Yes
6	Oval	Red	Yes
7	Round	Orange	No
8	Oval	Yellow	No
9	Round	Red	Yes

Classify a new fruit with the properties Shape = Oval and Color = Orange. Using the Naive Bayes algorithm, predict whether the fruit is edible or not.

- 9.a) Analyze the role of adjusting prior probabilities in handling class imbalance. How does this method affect model predictions, and in what types of classification problems would adjusting prior probabilities be most beneficial?

OR

- 9.b) Analyze the challenges of measuring predictor importance in the presence of severe class imbalance. How incorrect importance can measure skewed model interpretation, and what techniques can be employed to ensure accurate feature importance assessments?

10.a) Given the past values of a time series:

$$Y_{t-3}=50$$

$$Y_{t-2}=55$$

$$Y_{t-1}=60$$

Using an exponential smoothing model with a smoothing factor $\alpha=0.6$, forecast the next value Y_t . Discuss how the smoothing factor α influences the forecast.

OR

10.b) Given the time series data: [10, 15, 20, 25, 30] calculate the Lag-1 autocorrelation. What does the computed autocorrelation value indicate about the relationship between consecutive observations in the time series and discuss about the same?

↔↔↔ E/L/TX ↔↔↔