



FINAL EXAMINATION-2022

Course Title: Machine Learning

Course Code: CSE-4213

4th Year 2nd Semester

Session: 2019-20

Time: 3 hours

Marks: 60

Answer any five Questions from the followings.

1. a) Given a dataset of house prices with features (size, number of bedrooms, location), identify whether the problem is regression, classification, or clustering and justify. [3]
- b) Suppose you are building a model to predict loan approval. Describe the steps you would take from raw data collection to model deployment [5]
- c) Explain with an example how overfitting can occur and how it can be reduced. [4]

2. a) Given the following raw dataset, perform missing value handling and normalization for the "Age" and "Income" columns: [4]

Name	Age	Income (\$)	City
A	25	3500	Dhaka
B	NaN	2800	Chittagong
C	40	5000	Dhaka
D	30	NaN	Rajshahi

- b) Explain and demonstrate how label encoding can be applied to the "City" column. [4]
- c) Why is feature scaling important before applying KNN? Give an example. [4]

3. a) A classification model produces the following confusion matrix: [6]

	Predicted Positive	Predicted Negative
Actual Positive	50	10
Actual Negative	5	35

Compute Accuracy, Precision, Recall, and F1-Score.

- b) Briefly explain how cross-validation improves model reliability. [3]
- c) Differentiate between training accuracy and testing accuracy with example. [3]
4. a) Using the following data, classify a new instance using Naïve Bayes: [6]

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes

Find $P(\text{Play}=\text{Yes}|\text{Sunny})$ and $P(\text{Play}=\text{No}|\text{Sunny})$.

- b) Apply K-Means Clustering on the following 2D points for 2 clusters: (2,4), (3,5), (10,12), (11,11). Take initial centroids as (2,4) and (10,12). Perform one iteration and show cluster assignment. [6]
- a) Consider a set of five training examples given as $((x_i, y_i), c_i)$ values, where x_i and y_i are the two attribute values (positive integers) and c_i is the binary class label: $\{((1, 1), -1), ((1, 7), +1), ((3, 3), +1), ((5, 4), -1), ((2, 5), -1)\}$. [6]

Classify a test example at coordinates (3, 6) using a k-NN classifier with $k = 3$ and Manhattan distance defined by $d((u, v), (x, y)) = |u - x| + |v - y|$.
Onebyzero Edu - Organized Learning, Smooth Career
The Comprehensive Academic Study Platform for University Students in Bangladesh (www.onebyzerodedu.com) - 1.

- b) What value of k minimizes the training set error for this dataset? What is the resulting training error? [3]
 - c) Why might using too large values k be bad in this dataset? Why might too small values of k also be bad? [3]
6. a) Explain the basic idea of ensemble learning. Why do multiple weak learners often perform better together than a single strong model? [4]
- b) Differentiate between Bagging and Boosting in terms of their data sampling approach, error handling, and model combination strategy. Give one example algorithm for each. [8]
7. a) Explain the key components of a Convolutional Neural Network (CNN) such as convolutional layer, pooling layer, and fully connected layer. How do they work together in image classification? [4]
- b) A CNN is trained to detect vehicles from traffic images. Describe how feature extraction happens automatically through convolutional filters instead of manual feature design. [4]
- c) Discuss the role of overfitting in CNN models. Suggest two techniques (with explanation) to reduce overfitting in deep CNN architectures. [4]
8. a) Explain the concept of Linear Regression with an example. How is the relationship between dependent and independent variables represented mathematically? [4]
- b) Given a dataset of students' study hours (X) and their scores (Y), explain how the least squares method estimates the best-fit line. [4]
- c) Describe two key assumptions of linear regression. What problems can occur if these assumptions are violated? [4]