JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION- 2024

B.Tech-VII Semester (ECE)

COURSE CODE (CREDITS): 19B1WEC732 (3) MAX. MARKS: 35

COURSE NAME: Pattern Analysis in Machine Intelligence

COURSE INSTRUCTORS: Dr. Vikas Baghel MAX. TIME: 2 Hours

Note: *(a) All questions are compulsory.*

(b) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems

<table>
<tr><th>Q.No</th><th>Question</th><th>CO</th><th>Marks</th></tr>
<tr><td>Q1.</td><td>A company uses a pattern recognition system to classify fruits based on their size, weight, and color. The data for two fruits, apples and oranges, is represented as follows:
<table>
<tr><th>Fruit</th><th>Size (cm)</th><th>Weight (grams)</th><th>Color Code (Red: 1, Orange: 2)</th></tr>
<tr><td>Apple</td><td>7</td><td>150</td><td>1</td></tr>
<tr><td>Orange</td><td>8</td><td>170</td><td>2</td></tr>
</table>
i. Represent the fruits as patterns in a 3-dimensional feature space.
ii. Using the Euclidean distance, determine which fruit is closer to the query pattern (8 cm, 160 grams, color code: 2).
iii. Comment on the system's ability to generalize classification based on the provided data.</td><td>[CO1]</td><td>[1]
[2]
[2]</td></tr>
<tr><td>Q2.</td><td>A hospital is using a pattern recognition algorithm to classify patient data into two classes: diabetic (Class 1) and non-diabetic (Class 2). The feature vector includes fasting blood sugar (FBS, mg/dL) and HbA1c percentage (%). The training data for two patients is:
<table>
<tr><th>Patient</th><th>FBS (mg/dL)</th><th>HbA1c (%)</th><th>Class</th></tr>
<tr><td>A</td><td>126</td><td>6.5</td><td>1</td></tr>
<tr><td>B</td><td>100</td><td>5.5</td><td>2</td></tr>
</table>
i. Plot the feature space and label the two classes.
ii. A new patient has an FBS of 110 mg/dL and HbA1c of 6.0%. Using a simple distance-based classifier, determine the class of the new patient.
iii. Discuss how the choice of distance metric might affect classification in this scenario.</td><td>[CO2]</td><td>[1]
[2]
[2]</td></tr>
<tr><td>Q3.</td><td>A company is building a pattern recognition system to classify customer reviews as either 'positive' or 'negative' based on text features extracted from reviews. The company decides to use cosine similarity as a proximity measure for feature extraction. Consider the following term frequency-inverse document frequency (TF-IDF) vectors for two customer reviews:
• Review 1: (0.3, 0.4, 0.1, 0.2)
• Review 2: (0.5, 0.1, 0.3, 0.4)</td><td>[CO2]</td><td></td></tr>
</table>

i. Compute the cosine similarity between these two reviews. [3]

ii. Explain how proximity measures like cosine similarity help in pattern recognition, especially in text classification tasks. [2]

Q4. [CO3]

a) Discuss the role of the Linear Discriminant Function in solving classification problems in medical diagnosis. [1]

b) Illustrate how this supervised learning approach can be employed to classify patients based on disease severity using real-world data. [2]

c) Highlight the importance of feature selection and its impact on the performance of the classifier. [2]

Q5. [CO4]

a) Explain the concept of wrapper methods in machine learning for feature selection. Discuss how wrapper methods differ from filter and embedded methods. Highlight their advantages and disadvantages in the context of pattern classification tasks. [5]

b) A company is working on a project to reduce the file size of digital images by using hierarchical clustering for color quantization. Consider an image with 8 colors represented by the following RGB values (with each value in the range of 0 to 255):

Color	Red	Green	Blue
1	255	0	0
2	0	255	0
3	0	0	255
4	255	255	0

i. Apply Agglomerative Hierarchical Clustering with a Euclidean distance metric to merge the colors step by step. Construct the dendrogram for the clustering process. [3]

ii. Determine the number of clusters you would choose for the image compression and explain why. Based on the chosen clusters, compute the average RGB values for each cluster and suggest how the image can be quantized using the resultant colors. [2]

Q6. [CO5]

You are tasked with developing a spam email detection system using two classifiers: Support Vector Machine (SVM) and Naive Bayes (NB). Both classifiers are trained on the same dataset of emails, and their individual performance is as follows:

- SVM has an accuracy of 85%, precision of 80%, and recall of 90%.
- Naive Bayes (NB) has an accuracy of 82%, precision of 78%, and recall of 88%.

You decide to combine these classifiers using voting ensemble (majority voting) and stacking. In stacking, the predictions of the two models (SVM and NB) are used as input for a third model (logistic regression).

i. Calculate the combined accuracy, precision, and recall for the majority voting scheme if both classifiers classify an email as spam. [2]

ii. Explain the expected advantages and potential challenges of using stacking over majority voting in this scenario. [3]