JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST - 3 EXAMINATION - 2025

B.Tech. - VI Semester (BI)

COURSE CODE (CREDITS): 18B11BI611 (3) MAX. MARKS: 35

COURSE NAME: Machine Learning for Bioinformatics

COURSE INSTRUCTORS: D. Gupta MAX. TIME: 2 Hours

Note: *(a) All questions are compulsory. (b) The candidate can make suitable numeric assumptions wherever required to solve problems. (c)* ***Be concise****. (d) Use of a calculator is permitted.*

<table>
<tr><th>Q. No.</th><th>Question</th><th>CO</th><th>Marks</th></tr>
<tr><td>Q. 1</td><td>a) Explain how high bias causes underfitting and low bias causes overfitting in predicting student exam scores. Also, discuss how including relevant features like hours studied, attendance, and previous grades can help balance bias and improve the model's performance.

b) Briefly differentiate between stochastic, mini-batch, and batch gradient descent. Provide real-world examples to illustrate each method.

c) Explain the differences between Gini impurity and split information. Consider real-world examples to illustrate why these measures are important for making effective splits in machine learning applications.</td><td>1, 2, 4</td><td>[3]

[3]

[2]</td></tr>
<tr><td>Q. 2</td><td>a) Given the following observations from the survey with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Find the class of the test sample using k-NN algorithm. Take k=3 and use L2 norm for distance computations.

<table>
<tr><th>X1= Acid durability (seconds)</th><th>X2= Strength (kg/ square meter)</th><th>Y= Classification</th></tr>
<tr><td>7</td><td>7</td><td>Bad</td></tr>
<tr><td>7</td><td>4</td><td>Bad</td></tr>
<tr><td>3</td><td>4</td><td>Good</td></tr>
<tr><td>1</td><td>4</td><td>Good</td></tr>
</table>
Test Sample
<table>
<tr><td>3</td><td>7</td><td>?</td></tr>
</table></td><td>2, 3, 4</td><td>[4]</td></tr>
</table>

	b) What is the kernel trick in SVM? Explain how does the RBF kernel help in non-linear classification taking a real-world example.		[2]
Q. 3	a) Using the following dataset predict the class for the record (Confident=Yes, Sick=No) applying Naïve Bayes algorithm Confident | Studied | Sick | Result Yes | No | No | Fail Yes | No | Yes | Pass No | Yes | Yes | Fail No | Yes | No | Pass Yes | Yes | Yes | Pass	4	[5]
	b) Using the Laplace smoothing formula, explain the problem it solves in text classification with Naive Bayes and mention one limitation of this method?		[2]
Q. 4	a) Using the K-Means clustering algorithm with K=2, cluster the following data points: (2,3), (3,2), (4,4), (8,7), (7,8). Use Euclidean distance as the similarity measure. Initialize the algorithm by selecting the first two points as the initial centroids. Perform two iterations and clearly show the formation of clusters after each iteration.	5	[4]
	b) What is a dendrogram in hierarchical agglomerative clustering? Explain different linkage criteria affect the formation of clusters.		[3]
Q. 5	a) What is a genetic algorithm? Explain its main steps using the binary population: 1010, 1100, 1001, 0110 Perform one complete iteration of the algorithm - including fitness evaluation, selection, crossover (1-point), mutation (bit flip mutation), and show the resulting new population.	5	[4]
	b) Explain the Silhouette Coefficient and Dunn Index along with their formulas. How do these metrics help in determining the optimal number of clusters in clustering algorithms?		[3]