Dr. Ekta

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION- December, 2021

B.Tech. (CSE, IT) VII Semester

COURSE CODE: 19B1WCI731 MAX. MARKS: 35

COURSE NAME: Computational Data Analysis

COURSE CREDITS: 2 MAX. TIME: 2 Hrs.

Note: *All questions are compulsory. Carrying of mobile phone during examinations will be treated as case of unfair means.*

Q1 a. What are the key challenges in developing machine learning applications? [2]

b. Explain the principle of the gradient descent algorithm. Accompany your explanation with a labeled diagram. [2]

c. Differentiate between stochastic, batch and mini-batch gradient descent. [2]

Q2 a. What is overfitting and underfitting in machine learning. Express in terms of bias and variance. [2]

b. Using the following dataset, predict the class for the record ***(Color=Red, Type=SUV, Origin=Domestic)*** using Naïve Bayes algorithm. Explain all the steps clearly. [4]

Instance	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Q3 a. What is a dendrogram in hierarchical clustering? How to get the optimal number of clusters using a dendrogram? [3]

b. Consider the following 8 data points with (x, y) representing locations. Use k-means clustering algorithm to group these into three clusters. [3]

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Note: Consider the initial cluster centers as **A1(2, 10), A4(5, 8) and A7(1, 2).** The distance function between two data points a = (x1, y1) and b = (x2, y2) is defined as:

$$P(a, b) = |x2 - x1| + |y2 - y1|$$

P.T.O

Q4 a. Explain Adaboost algorithm with the help of an example. [3]

b. List at least four differences between bagging and boosting ensemble learning techniques. [3]

Q5 a. What are the objectives of feature selection methods? [2]

b. Consider the following set of training examples: [3]

Instance	Classification	F1	F2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

What is the information gain of F2 relative to these training examples? Write the equation for calculating the information gain and intermediate results.

Q6 a. Mohit built a logistic regression model with a training accuracy of 97% and a test accuracy of 51%. What could be the possible reasons for the gap between these accuracies? How this problem can be solved? [3]

b. List at least four differences between L1 and L2 regularization in regression. [3]