Charotar University of science & Technology (CHARUSAT)
Devang Patel Institute of Advance Technology and Research
Unit test 1

Subject Name: Machine Learning (CS344)
Date: 17/01/2020
Time: 9:45 to 10:45
Total Marks: 30

SET 1

1. In which algorithm computation time is required more to test unseen samples? **(1 Mark)**
   a. K-Nearest Neighbors
   b. Decision Tree
   c. SVM
   d. Neighborhood

2. Cluster quality depends on _____ intra-class distance and _____ inter-class distance. **(1 Mark)**
   a. average, minimum
   b. minimum, maximum
   c. maximum, minimum
   d. minimum, average

3. The K-means clustering algorithm is not sensitive to outliers. **(1 Mark)**
   a. True
   b. False

4. Following is not a library of Deep Learning Algorithm. **(1 Mark)**
   a. Tensorflow
   b. PyTorch
   c. Keras
   d. Sci-Kit Learn

5. import pandas as pdo

   dfo = pdo.read_csv("autom.csv")

   dfo = dfo [['company','price']][dfo.price==dfo['price'].max()]

   dfo

index	company	body-style	wheel-base	length	engine-type	num_cylinders	hp	avg_mileage	price
0	alpha-romero	convertible	88.6	168.8	dohc	four	111	21	13495.0
1	alpha-romero	convertible	88.6	168.8	dohc	four	111	21	16500.0
2	alpha-romero	hatchback	94.5	171.2	ohcv	six	154	19	16500.0
3	audi	sedan	99.8	176.6	ohc	four	102	24	13950.0

4	audi	sedan	99.4	176.6	ohc	five	115	18	17450.0

What is the outcome of above query? **(1 Mark)**

a. the most expensive car price
b. the most expensive car company name
c. the most expensive car company name with price
d. None of the above

6. Gini Index would be _____ if dataset is perfectly classified. **(1 Mark)**
   a. 0
   b. 1
   c. 1/2
   d. 1/3

7. If the samples are an equally divided by target classes, it has entropy of _____. **(1 Mark)**
   a. 0
   b. 1
   c. 1/2
   d. 1/3

8. In machine learning, most of the applied features need to be identified by an expert before feeding to an algorithm compared to deep learning. **(1 Mark)**
   a. True
   b. False

9. If the data is skewed, _____ is a better measure of central tendency. **(1 Mark)**
   a. Mean
   b. Mode
   c. median
   d. none of the above

10. How do you handle missing or corrupted data in a dataset? **(1 Mark)**
    a. Drop missing rows or columns
    b. Replace missing values with mean/median/mode
    c. Assign a unique category to missing values
    d. All of the above

11. Consider a linear-regression model, Y = mX + C. Values of X = [2,4,5,6] and Y = [4,6,7,8]. Calculate MSE loss. Take, m = 0.5 and b = 0.2 **(3 Mark)**
    a. 15.95
    b. 63.81
    c. 95.15
    d. 23.04

12. Find the variance for the following set of data representing trees in California (heights in feet): 6, 21, 98, 200, 18, 10 **(2 Mark)**
    a. 5744.36
    b. 5447.63
    c. 5497.63
    d. 5947.36

13. For the below data, where 'Eat Pizza?' is the target variable, what will be the root node using Information Gain? **(5 Mark)**

Time of Day	Day	No of People	Income	Eat Pizza?
Day	Weekday	4	Medium	No
Night	Weekend	3	Low	No
Day	Weekday	2	High	No
Day	Weekday	4	Medium	No
Day	Weekday	4	Low	No
Night	Weekend	1	High	Yes
Day	Weekend	2	High	Yes
Night	Weekend	3	Medium	Yes
Day	Weekday	4	Low	No
Night	Weekend	3	Low	Yes
Night	Weekend	2	Medium	Yes
Day	Weekend	1	Low	Yes
Night	Weekday	2	High	Yes
Night	Weekend	3	Medium	Yes

a. Time of Day
b. Day
c. Income
d. No of People

14. For k = 3, and Centers initialized as C1 = P1, C2 = P2, C3 = P3, what will be the Clusters after the first iteration of k-means clustering algorithm? Use Manhattan distance instead of Euclidean distance. Manhattan Distance for two 2D points i and j = $(|x_i - x_j| + |y_i - y_j|)$
(5 Mark)

	X	Y	Z
P1	**2**	**5**	**1**
P2	**3**	**3**	**4**
P3	**5**	**4**	**3**
P4	**5**	**7**	**4**
P5	**4**	**5**	**1**
P6	**1**	**1**	**6**
P7	**1**	**6**	**1**

P8	**4**	**2**	**8**

a. {P1,P2,P3}, {P4,P5}, {P6,P7,P8}
b. {P1, P5, P7}, {P2, P6}, {P3, P4, P8}
c. {P1}, {P2, P4, P5, P6, P7, P8}, {P3}
d. {P1, P5, P7}, {P2, P6, P8}, {P3, P4}

15. Using K-Nearest Neighbors, what will be the values marked as "?". k = 5. Raining = 1 indicates that it is raining and 0 indicates that it is not raining.
Use Manhattan distance as a measure. Manhattan distance of two points i and j = $|x_i-x_j| + |y_i-y_j|$ **(5 Mark)**

ID	Temperature	Wind Speed	Raining
1	**5**	**0.4**	**1**
2	**17**	**1.5**	**0**
3	**7**	**5.0**	**1**
4	**10**	**3.5**	**1**
5	**22**	**2.2**	**0**
6	**13**	**4.5**	**1**
7	**15**	**3.0**	**1**
8	**25**	**2.6**	**0**
9	**20**	**1.0**	**1**
10	**30**	**5.6**	**0**
11	**18**	**3.2**	**?**
12	**20**	**2.0**	**?**

a. 0, 0
b. 0, 1
c. 1, 0
d. 1, 1