

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 2**

1. 

```
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
- 
2. Following is not a library of Deep Learning Algorithms **(1 mark)**
    - a. Tensorflow
    - b. PyTorch
    - c. Keras
    - d. Sci-Kit Learn
- 
3. 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
- 
4. 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
5. The K-means clustering algorithm is not sensitive to outliers. **(1 mark)**  
a. True  
b. False
6. Gini Index would be \_\_\_\_\_ if dataset is perfectly classified. **(1 mark)**  
a. 0  
b. 1  
c.  $\frac{1}{2}$   
d.  $\frac{1}{3}$
7. 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
8. 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
9. If the samples are an equally divided by target classes, it has entropy of \_\_\_\_\_. **(1 mark)**  
a. 0  
b. 1  
c.  $\frac{1}{2}$   
d.  $\frac{1}{3}$
10. 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
11. 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
12. 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
13. For the below data, where 'Eat Pizza?' is the target variable, which feature has the Lowest Gini Index? **(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

- Day
  - Time of Day
  - No of People
  - Income
14. For  $k = 3$ , and Centers initialized as  $C1 = P4$ ,  $C2 = P5$ ,  $C3 = P6$ , what will be the Cluster Centers 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	1	4	0
P2	2	7	3
P3	4	3	1
P4	4	6	3
P5	3	4	0
P6	0	0	5
P7	0	5	0
P8	3	1	7

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

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	0
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	0
10	30	5.6	0
11	21	4.7	?
12	11	2.6	?

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