National Institute of Technology
Tiruchirappalli, Tamil Nadu – 620 015

Machine Learning for Engineering Applications – CT2 **Date: 01.05.2021**

Duration: 1 Hr **Time:** 10:40 – 11:40 AM

Total Marks: 20

Note: *Some MCQs may have multiple answers. In such case, you have to write all the correct choices. Otherwise, no marks will be provided for that question.*

1. What information does one understand by applying covariance between two features? ***(1 M)***

   (a) Only the direction of relationship **(b)** Only the strength of relationship

   (c) Both direction and strength of relationship **(d)** None of the above

2. Assume that you are planning to apply Decision Tree algorithm and your dataset is not having any outlier. But the scale/range of values in columns and the units of columns are different. In such case, do you have to do Feature Scaling? ***(1 M)***

3. (i) Does the feature "Age" in below dataset require Feature Scaling? State the reason.

   (ii) How will you identify whether the features "Speed" and "Acceleration" are related to each other or not (State the reason). Also, draw a rough graph and identify what will be the range/value of Pearson Correlation Coefficient. What will be your final decision – Can we drop any one of these feature or not? ***(1 + 4 = 5 M)***

Sl. No.	Name	Age	Speed (in km/hr)	Acceleration (in %)	Rotation of Tyre (in km/hr)	Direction of Motion	Clear Vision (in %)	Class
1.	Bala	1000	100	60	100	Forward	80	Racing
2.	Karthick	25	120	70	120	Forward	70	Racing
3.	Sundar	35	140	75	140	Forward	60	Racing
4.	Rajesh	30	160	85	160	Forward	50	Racing
5.	Kumar	20	180	100	180	Forward	40	Racing

4. Which method can detect non-linear correlation between features? ***(1 M)***

   (a) Pearson correlation **(b)** Spearman Correlation **(c)** None of the above

5. What is the aim of Principal Component Analysis? ***(1 M)***

   (a) Normalize the feature values
   (b) Facilitate to plot the dataset in a lesser dimensional space
   (c) Facilitate to identify the direction of spread of data/information and include only these axises in the plot
   (d) None of the above

6. If a model performs well for training dataset and produces a lot of errors during testing time, then the model is said to be _____ ***(1 M)***

   (a) Underfitting **(b)** Overfitting **(c)** Perfect fit **(d)** Having High Bias and Low Variance **(e)** Having Low Bias and High Variance

7. Consider that I have designed a classification model that aims to detect criminal's records. The sample dataset has both criminal's and general public's records. The developed model has the following confusion matrix. What do you understand from it?

   *[**Hint:** Don't apply any formula. Just discuss in terms of: (i) No. of classes; (ii) Total no. of samples; (iii) No. of Samples in each class; (iv) No. of TP, TN, FP, FN; (v) Whether the developed model is a good one or not – State the reason]* ***(4 M)***

	Predicted Class		
Actual Class		**Public**	**Criminal**
	Criminal	10	113
	Public	60	8

8. Does the Principal Component Analysis sensitive to relative scaling? ***(1 M)***

   (a) Yes **(b)** No

9. Consider that I have developed a classification model that aims to identify apples from the given samples. Suppose, my dataset has 100 apples and 50 oranges. My model is able to identify 80 apples and 40 oranges correctly. Draw the confusion matrix for the above information and also specify TP, TN, FP and FN. ***(3 M)***

10. Which of the following information is correct about ROC curve? ***(1 M)***

    (a) False Positive Rate vs True Positive Rate
    (b) Precision vs Recall
    (c) Recall vs False Positive Rate
    (d) Specificity vs Sensitivity

11. What does the term “k” signify in k-Means Algorithm? ***(1 M)***

(a) Only the no. of samples that has to be considered to form a cluster
(b) Only the no. of clusters to be formed
(c) Both (a) and (b)
(d) None of the above

---------- END ----------

Date

Satyam Singh
11211906
Sub:- CSOE 18
CT-2

Sol1. (a) only the direction of relationship.

Sol5. (b) Facilitate to plot the dataset in lesser dimensional space.
(c) Facilitate to identify the direction of spread of data / information and include only these axises in the plot

Sol6. (b) overfitting
(c) Having low Bias and high variance

Sol4. (b) Spearman correlation

Sol10 (c) Recall vs False positive Rate

Sol11 (b) only the no. of clusters to be formed.

Sol8. (a) Yes

Sol2. No

Sol9. No. of apple = 100
No. of orange = 50

		Actual class	
Predicted class	~~actual~~	apple	orange
	apple	80	10
	orange	20	40

TP = 80 FN = 20
TN = 40 FP = 10

Satyam Singh
112119066

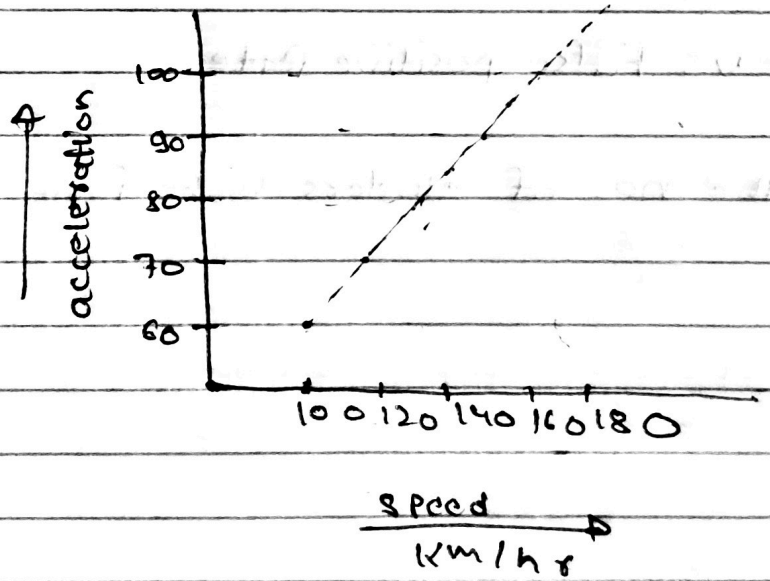
Date

Sol 3.(i) Yes. in this dataset it requires feature scaling, because '1000' value is there by mistake, else usually it's not necessary.

(ii) * Pearson's correlation:- Helps to identify if 'speed' and 'acceleration' are related or not, this is done so that, of the column could be dropped if the correlation is high, which can be found by the formula.

$$P_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Plot



* value of pearson correlation coefficient $\Rightarrow \rho = 1$, because all the value are on the line.

* yes, we should drop one of the feature.

Satyam Singh - 11211906 6

Date

Sol 7 The clarification model to detect true criminal record is

		Predicted class	
Actual class		Public	Criminal
	Criminal	10	113
	Public	60	8

⇓

Convert this in the form of

		Actual class	
Predicted Class		Criminal	Public
	Criminal	113	8
	Public	10	60

So,

(i) No. of classes = 2

(ii) Total no. of sample = 8 + 60 + 113 + 10
= 191

(iii) Sample in public = 68
Sample in criminal class = 113

(iv) Since criminal is positive
TP = 113 , TN = 60
FN = 10 , FP = 8

(v) accuracy = $\frac{TP + TN}{\text{Total}}$

$= \frac{113 + 60}{191}$

$= 90.575$

which implies that it is a good model.