National Institute of Technology
Tiruchirappalli, Tamil Nadu – 620 015

Machine Learning for Engineering Applications – FA **Date: 18.05.2021**

Duration: 1 Hr 30 Min **Time:** 03:00 – 04:30 PM

Total Marks: 30

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. Name the three ways by which the ensemble models try to consolidate the predicted outcomes. ***(2 M)***

2. What kind of information does the dendrogram signify? ***(2 M)***

   (a) Only the order of the cluster formed
   (b) Only the similarity of the cluster
   (c) Both similarity and order of the clusters formed
   (d) None of the above

3. (i) Consider the following dataset where y is the actual value and y' is the predicted value. Find the value of R^2. Also, state whether the model is good or not.
   *[**Hint:** If the value of R^2 is greater than 1, then consider the value as 1]* ***(6 M + 1.5M= 7.5 M)***

X	Y	Y'
1	11	11
2	4	3.8
3	6	5.6
4	9	9.4
5	2	2.5

Dr. R. Bala Krishnan, Assistant Professor, Department of CSE

4. Calculate the spearman correlation coefficient value for the following dataset. State what do you observe from the obtained value. ***(6 M + 1.5 M = 7.5 M)***

X	Y
40	29
150	25
135	23
85	53
95	15
102	96

5. Write the name of the normalization/standardization technique that makes the data to follow the normal distribution. ***(2 M)***

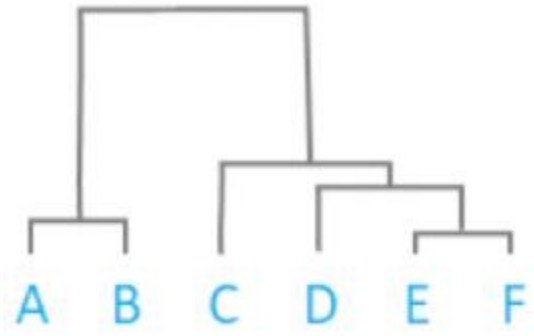
6. In linear regression, the difference between the actual and predicted values are called as_____

   (a) Difference **(b)** Residual error **(c)** Intercept **(d)** Slope

   (2 M)

7. Draw box-and-whisker plot for the following values: **19, 11, 9, 8, 4, 50, 15** ***(5 M)***

8. For the following dendrogram, write the order in which the clusters are being formed.
   *[**Hint:** Draw the diagram and give numbers starting from 1 to n].* ***(2 M)***



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

National Institute of Technology
Tiruchirappalli, Tamil Nadu – 620 015

Machine Learning for Engineering Applications – FA **Date: 18.05.2021**

Duration: 1 Hr 30 Min **Time:** 03:00 – 04:30 PM

Total Marks: 30

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. Which of the following information is correct about stratified cross-validation ***(2 M)***
   (a) Splits the data into "k" random folds
   (b) Split the data into "k" folds ensuring that the number of representation from each class is similar in every fold
   (c) Both (a) and (b)
   (d) None of the above

2. Consider the following data and apply smooth by bin means and smooth by bin boundary techniques *[**Hint:** Consider, Bin Size = 3]* ***(6 M)***
   25, 21, 95, 6, 53, 17, 45, 61, 12, 4, 11, 5

3. If you build a decision tree model to it's complete depth, then what problem does the model face (if any)? ***(2 M)***
   (a) Over Fitting **(b)** Under Fitting **(c)** Both (a) and (b) **(d)** None of the above

4. Suppose your model has predicted that the provided sample belongs to the negative class whereas the sample actually belongs to the positive class. Then the terminology used to represent this condition is called as _____ ***(2 M)***
   (a) True Positive **(b)** True Negative **(c)** False Positive **(d)** False Negative

5. Assume that you are going to build a random forest comprising of three models internally. Now, I want to use the following dataset to train my models. Identify the possible training datasets for each model by considering both row and feature sampling with replacement technique. ***(7 M)***

Shape	Diag	# of Diag	Length (in cm)	Height (in cm)	Target Class
Rectangle	Yes	2	100	10	1
Rectangle	Yes	2	1000	100	1
Square	Yes	2	100	100	0
Square	No	2	10000	10000	0
Triangle	No	0	100	1000	0
Triangle	Yes	0	50	1000	0
Circle	No	0	100	100	1
Circle	Yes	0	50	50	1
Diamond	Yes	2	100	100	0
Rectangle	Yes	2	100	5	1

Train Dataset

Test Dataset

6. The name of the method which builds multiple models and then consolidate the predictions made by the individual models in order to take the final decision is called as_____ ***(2 M)***

(a) Supervised **(b)** Unsupervised **(c)** Ensemble **(d)** Reinforcement

7. Calculate the spearman correlation coefficient value for the following dataset. State what you observe from the obtained value. ***(6 M + 1 M = 7 M)***

X	Y
95	30
131	45
165	56
75	11
35	28
149	7

8. If the value of covariance between two features is **positive** and **< 0.3**, then what does it signify?

(2 M)

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

National Institute of Technology
Tiruchirappalli, Tamil Nadu – 620 015

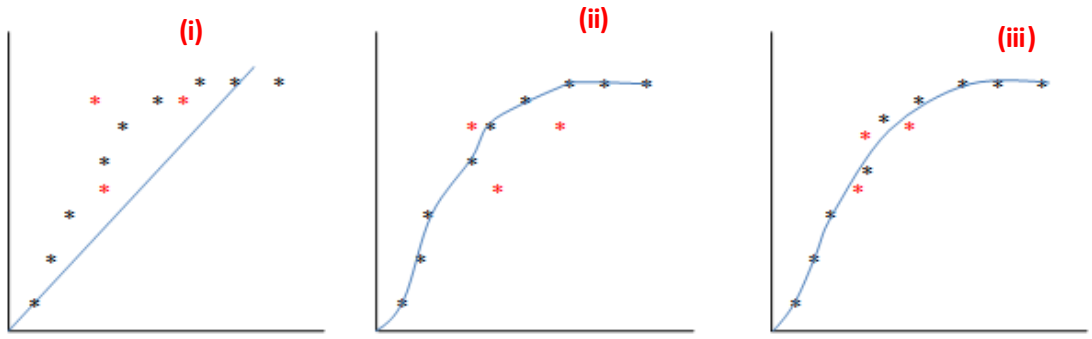
Machine Learning for Engineering Applications – FA **Date: 18.05.2021**

Duration: 1 Hr 30 Min **Time:** 03:00 – 04:30 PM

Total Marks: 30

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. Match the following plot with the corresponding terminologies *[**Hint:** Consider that the black asterisk represents the training samples and red asterisk represents the test samples].* ***(1 M)***



(a) (i) Perfect Fit; (ii) Under Fitting; (iii) Over Fitting
(b) (i) Over Fitting; (ii) Perfect Fitting; (iii) Under Fitting
(c) (i) Under Fitting; (ii) Perfect Fitting; (iii) Over Fitting
(d) (i) Under Fitting; (ii) Over Fitting; (iii) Perfect Fitting

2. Draw box-and-whisker plot for the following values: **29, 10, 9, 12, 4, 50, 15** ***(5 M)***

3. Hierarchical clustering tries to ———————— ***(1 M)***
   (a) Put the data into the number of clusters you tell it to
   (b) Tell you what two things are pair-wise similar
   (c) Both (a) and (b)
   (d) None of the above

4. Consider that I have developed a classification model that aims to identify the vehicles that have met with an accident from the given samples. Suppose, my dataset has 500 non-accident vehicles and 100 accident vehicles. My model is able to identify all the 500 non-accident vehicles and only 10 accident vehicles correctly. ***(4 M + 1 M + 1M=6 M)***
   (i) Draw the confusion matrix for the above information
   (ii) State whether the developed model is a good one or not
   (iii) State whether any bias exists in the model. If so, state the bias exists in predicting which class?

5. Consider the following dataset where y is the actual value and y' is the predicted value. Find the value of R^2. Also, state whether the model is good or not.
   *[**Hint:** If the value of R^2 is greater than 1, then consider the value as 1]* ***(6 M + 1.5M= 7.5 M)***

X	Y	Y'
1	4	3.6
2	3	3.2
3	1	0.8
4	3	2.5
5	5	4

6. Calculate the spearman correlation coefficient value for the following dataset. State what you observe from the obtained value. ***(6 + 1.5 M = 7.5 M)***

X	Y
120	35
151	43
140	65
65	14
85	11
110	9

7. Write the formula for Precision in terms of TP, TN, FP, FN ***(2 M)***

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

National Institute of Technology
Tiruchirappalli, Tamil Nadu – 620 015

Machine Learning for Engineering Applications – FA **Date: 18.05.2021**

Duration: 1 Hr 30 Min **Time:** 03:00 – 04:30 PM

Total Marks: 30

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. Write the names of the two distance measure that are used to evaluate the distance between the data points in K-Means algorithm. ***(2 M)***

2. Suppose that you are applying Pearson Correlation Coefficient technique on a dataset comprising of 4 features namely A, B, C and D. You got the correlation values between A and B as -0.875, A and C as 0.95, A and D as 0.025, B and C as -0.46, c and d as 0.96, B and D as 0.6. Represent the correlation values in a matrix form *[**Hint:** Write the whole matrix with symmetry structure]*. ***(5 M)***

3. Suppose that your model has predicted that the provided sample belongs to the positive class and the sample actually belongs to the positive class. Then the terminology used to represent this condition is called as _____ ***(2 M)***

   (a) True Positive **(b)** True Negative **(c)** False Positive **(d)** False Negative

4. Which of the following is correct? ***(2 M)***

   (a) Pearson Correlation – Linear relationship; Spearman Correlation – Linear Relationship
   (b) Pearson Correlation – Non-Linear relationship; Spearman Correlation – Non-Linear Relationship
   (c) Pearson Correlation – Non-Linear relationship; Spearman Correlation – Linear Relationship
   (d) Pearson Correlation – Linear relationship; Spearman Correlation – Non-Linear Relationship

5. (i) Does dimensionality reduction and feature selection technique mean the same? ***(2 M)***
   (a) Yes **(b)** No

Dr. R. Bala Krishnan, Assistant Professor, Department of CSE

6. Consider the following dataset where y is the actual value and y' is the predicted value. Find the value of R^2. Also, state whether the model is good or not.
   *[**Hint:** If the value of R^2 is greater than 1, then consider the value as 1]* ***(6 M + 1.5M= 7.5 M)***

X	Y	Y'
1	9	8.9
2	3	2.5
3	6	5.6
4	7	6.4
5	2	1.8

7. Calculate the spearman correlation coefficient value for the following dataset. State what you observe from the obtained value. ***(6 M + 1.5 M = 7.5 M)***

X	Y
100	29
150	35
120	23
45	11
95	15
102	19

8. Write the formula for Recall in terms of TP, TN, FP, FN ***(2 M)***

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