# END SEM , ME-781, November 21, 2022

Max Marks: 200, Total time: 3 Hr

- No explanation for any question would be provided.
- Please make any assumptions as you see fit and solve the questions.
- This is an open-notes exam.
- You need not derive anything from scratch if it was derived in the class.
- You are not allowed to use a computer or calculator.
- You are not required to numerically solve an arithmetic expression.

5
10
10

1.
   a. For a simple linear regression, show that the least-squares line always passes through the point $(\bar{x}, \bar{y})$.
   b. For a generalized multiple linear regression show that $(y - \hat{y})$ is orthogonal to $\hat{y}$.
   c. Is there a relationship between $\sum_1^n (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^2$ , $\sum_1^n (y_{mean} - \boldsymbol{\hat{y}_i})^2$ and $\sum_1^n (y_i - y_{mean})^2$, where $\boldsymbol{y_i}$ is the response variable in the data set, $\boldsymbol{\hat{y}_i}$ in the predicted value and $y_{mean}$ is the mean of $\boldsymbol{y_i}$ in the dataset. The entire dataset is used for both training and validation.

10
10

2. An insurance company developed a Linear Discriminant Analysis classifier to predict if a person would need a reading glass or not based on the person's age. A dataset of 1000 persons is collected for this model development. If the average age of persons without glass was 30 years and that with the glass was 50 years, and 500 persons from the data set needed glasses, and the other 500 did not, find the Bayes decision boundary for this classifier. What would this decision boundary be if there were 600 people in the data set who needed glasses? Write the algebraic expressions without numerically solving them. Assume that the age of persons in each dataset is normally distributed and has the same variance.

5
5

3. In a supervised learning model, briefly explain how to handle:
   a. when the predictor variable is of nominal data type
   b. when the output (the response) variable is of nominal data type

5

4. In a random household with two children, let the event that the first child is a boy be ***A***, and the event that both the children are boys be ***B***. Using conditional probability, find out whether these two events are independent or not.

5
5

5. The eigenvalues and eigenvectors for a matrix $C = \begin{bmatrix} 1 & 2 & 1 \\ 6 & -1 & 0 \\ -1 & -2 & -1 \end{bmatrix}$.

   are as follows:

   Eigenvalues $\lambda 1 = 0$, $\lambda 2 = 3$ and $\lambda 3 = -4$

Eigenvectors $a1 = \begin{bmatrix} -\frac{1}{13} \\ -\frac{6}{13} \\ 1 \end{bmatrix}$, $a2 = \begin{bmatrix} -1 \\ -\frac{3}{2} \\ 1 \end{bmatrix}$ and $a3 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$

What would be the eigenvalues and eigenvectors for the following matrix :

a. $C1 = \begin{bmatrix} 3 & 2 & 1 \\ 6 & 1 & 0 \\ -1 & -2 & 1 \end{bmatrix}$.  b. $C2 = C + C^2$

5
10

5. Briefly explain:
   a. In simple linear regression, what does a low p-value for t-statistic indicate?
   b. For multiple linear regression with a large number of predictor variables, why t-statistic for each variable cannot be used as an argument to reject the null hypothesis?

5

6. For which of the following two cases of multiple linear regression is the null hypothesis more likely to be true:
   - F=1.1 for the F-statistic and number of data points, n= 500, and the number of predictor variables are 25
   - F=1.1 for the F-statistic and number of data points, n= 500, and the number of predictor variables are 20

10

7. A measure to assess the goodness of a machine learning model is to compare the mean squared error of testing. Without testing, an estimate of the mean square error for testing can be performed using Cp and BIC (on the training). Show that for any reasonably sized training dataset, BIC is a more stringent criterion. What could be the size of a reasonably sized dataset?

10

8. Graphically show that in Lasso regression (for a linear model with two predictor variables), reduction of a Coefficient (of a predictor variable) to zero value, with an increase in the penalty term, necessarily means that it will remain 0 on further increase in the penalty term.

10
10

9. Briefly explain:
   a. How will the mean squared error of training and testing change with overfitting? How will the amount of overfitting change by increasing the size of the dataset?
   b. How will model bias and model variance change with an increase in flexibility?

5

10. In a computer chip manufacturing company, the possibility of a manufacturing defect is 0.1%. A classification model is developed to predict defective parts (classify the product in one of the two classes: Defective, Non-defective) based on the process parameters. The model predicts non-defective parts correctly 99.9% of the time. Is this a good classification model? Briefly explain and provide an alternative methodology to assess the goodness of the model.

5
5

11. Very briefly explain one of the biggest strengths and the biggest weaknesses of:
    a. K-nearest neighbor classification model
    b. Random Forests classification

5
5

12. What would be the difference in a Logistic regression model for a response variable of many classes when the response variable is of nominal data type and when it is of ordinal data type? Briefly explain.

5

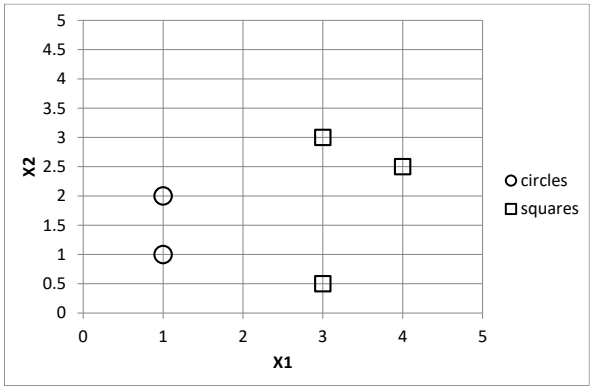
13. In K-Means Clustering Algorithm, there could be many different possible cluster formations depending upon the initial random assignment. What could be possible measures to reduce this problem?

15

14. In a k-Fold Cross-Validation process of 5N data points, the data set is randomly divided into 5-folds (5 equal groups) of N data points each. Four of the five folds are used for training, and the remaining one fold is taken for testing (validation). In this process, how much more computational effort would be required as compared to the Leave-One-Out Cross-Validation strategy if we want to try all possible combinations in Leave-One-Out Cross-Validation and all possible combinations in k-Fold Cross-Validation (for the same initial, randomly divided, 5-folds data set). Assume that the computational efforts are only required for model training and are given by T(n), where n is the number of data points in the training set.

10
10

15. A data set of circles and squares in two-dimensional space is provided below. A classification model needs to be developed which can identify circles from squares for an unknown point in the 2D space. Schematically draw the first nearest neighbor decision boundary and the linear maximal margin classifier decision boundary for the given data set.




How would these boundaries change (show by drawing) if one more circle data point is present at (1,4).

5

16. In a ML modeling exercise, the response variable is not a function of a few of the predictor variables. Therefore, to reduce the number of predictor variables, should one apply PCA or subset-selection strategy? Briefly explain.