Group 1

Name: **Reg Num**

1. (1 mark) Which of the following imputation is most appropriate for a categorical feature?
(a) Mean Imputation
(b) Grouped Mean Imputation
(c) Median Imputation
(d) Mode Imputation

2. (1 mark) Logistic Regression is used for
(a) binary classification
(b) multiclass classification
(c) both

3. (1 mark) Log normal distributions are
(a) left skewed
(b) heavily left skewed
(c) heavily right skewed
(d) symmetric

4 (1 mark) Which of the following is equivalent to Within Cluster Sum of Squares (WCSS) value for a given cluster?
(a) Cluster centroid
(b) Cluster variance
(c) Cluster median
(d) Cluster standard deviation
(e) Cluster mean absolute deviation

5. (1 mark) A min max scaler is given by (x - xmin)/(xmax - xmin). What will be the range of this new feature?
[0,1]

6. (1 mark) Which of these is least sensitive to outliers?
1. Mean
2. **Median**
3. Standard deviation

7. (4 marks) A toy dataset D = {(-1, 3) (-1, 2), (1,4) (2,5) } is provided. Assume k = 2 and perform KMeans clustering for 1 iteration using Expectation Maximization algorithm. Choose (-1,3) and (2, 5) as the initial random centroids.

Group 2

Name: **Reg Num**

1. (1 mark) If N is the size of the dataset, then selecting K = 1 in KNN causes
(a) Curse of dimensionality
(b) underfitting
(c) overfitting

2. (1 mark) A dataset has a column called “Country”, is categorical feature & takes values {India, Pakistan, Srilanka} Which of the following encoding is best?
(a) Label Encoding
(b) One Hot Encoding
(c) Ordinal Encoding
(d) Binary Encoding
(e) Factor Encoding

3. (1 mark) For your chosen answer in the previous question, perform that encoding on the feature and show all possible values

India	0	0	1
Pakistan	0	1	0
Srilanka	1	0	0

4. (1 mark) Fill in the blanks: Basic Nearest Centroid Model takes into account __________________
(a) both mean and variance
(b) mean, but not variance
(c) centroid and variance
(d) variance, but not mean

5. (1 mark) A min max scaler is given by (x - xmin)/(xmax - xmin). What will be the range of this new feature?
[0,1]

6. (1 mark) Why is k chosen as an odd number in kNN? (1 sentence answer)

Even number has the potential to produce tie. Odd number does not have that issue

7. (4 marks) A toy dataset D = {(-1, 3) (-1, 2), (1,4) (2,5) } is provided. Assume k = 2 and perform KMeans clustering for 1 iteration using Expectation Maximization algorithm. Choose (-1,3) and (2, 5) as the initial random centroids.

Group 3

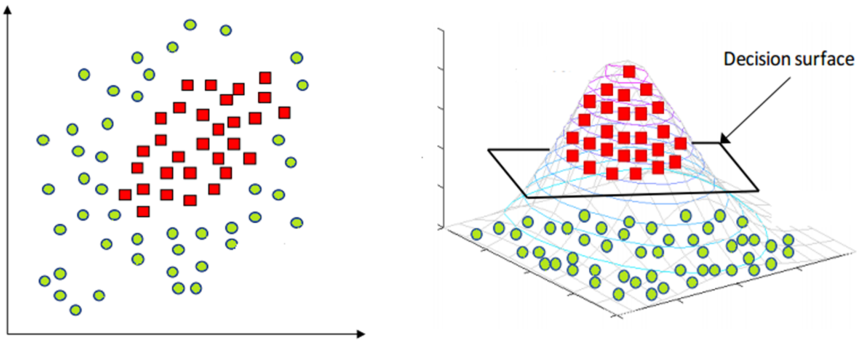
Name: **Reg Num**

1. (1 mark) Identify the scenario where clustering is not most appropriate
(a) Customer Segmentation when group identifier is not known
(b) Identifying outliers when the class membership anomaly/normal is unknown
(c) Identifying the groups in Iris dataset when labels are given

2. (2 marks) You are given a dataset with million records. You are asked to perform regression. Will you choose Polynomial Regression or neural networks based non-linear regression? When will you choose the other option? Give only a compact 1-2 sentence reason for choosing one against another.

A polynomial regression, while better than linear regression still has a fixed structure, to which the data may not adhere. When n > 1 million, choosing a Neural Networks regressor will definitely give a good model without the risk of overfitting. The downside of Neural Networks is that GPU is needed, training cycles may be long.

3. (2 mark) How will you convert a classification problem with inherently non linear decision boundary into a classification problem with linear decision boundary? Answer in 1-2 compact sentences at max. You can draw a diagram if you wish



4. (1 mark) kNN is trained for hyperparameters k = {3, 5, 7}, distance = {"manhattan", "euclidean"} and weight = {"uniform", "distance"} How many times does the model get trained in total when GridSearch with KFold CV=3 is performed over the hyperparameters?
3 x 2 x 2 x 3 + 1 = 37 times

5. (1 mark) RobustScaler is given by (x-q2)/(q3-q1). Why is it robust to outliers? 1 sentence answer
Because, Median and IQR which respectively make the numerator and denominator both are robust to outliers

6. (1 mark) Why would you like to have a feature as normal distributions in a dataset? 1-2 sentence answer

   a. **Faster and stable convergence.**
   b. **Z transforms make most sense for Gaussian**
   c. **Uncorrelated features in Gaussian imply independent feature only in Gaussian**

7. (2 marks) Draw the rough sketch of bivariate gaussian distributions and their decision boundary for a dataset with 2 classes. Gaussian distribution for both classes have the same covariance matrix and same variance for both features, but the proportion of the data belonging to class 1 is more than proportion of data belonging to class 2