**2 marks question:**

How do you handle missing or corrupted data in a dataset?

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**

What is the purpose of performing cross validation?

1. To asses the predictive performance of the models
2. To judge how the trained model performs outside the sample on test data.
3. **Both A & B**

Why is second order differencing in time series needed?

1. To remove stationarity
2. To find the maxima or minima at the local point
3. **Both A & B**

When performing regression or classification which of the following is the correct way to pre-process the data?

 **Normalize the data > PCA (Principal Component Analysis) > Training**

Which of the following Is an example of feature extraction?

1. Constructing bag of words vector from an email
2. Applying PCA projects to a large high dimensional data
3. Removing stopwards in a sentence
4. **All of the above**

Which of the following is true about Naïve bay’s algorithm?

1. Assume that all the features in a data set are equally important
2. Assume that all the features in a data set are independent
3. **Both A & B**

Which of the following statements about regularisation is not correct?

1. Using too a large a value of lamda can cause your hypothesis to underfit the data
2. Using too a large a value of lamda can cause your hypothesis to overfit the data
3. Using a very large value of lamda cannot hurt the performance of your hypothesis
4. **None of the above**

How can you prevent a clustering algorithm from getting stuck in bad local optima?

1. Set the same seed value for each run
2. Use multiple random initializations
3. **Both A & B**

Which of the following techniques can be used for normalization in text mining?

1. Stemming
2. Lemmatization
3. Stopward removal
4. **Both A & B**

In which of the following cases will K means clustering fail to give good results?

1. Data points with outliers
2. Data points with different densities
3. Data points with non convex shapes

**For all the three cases**

What is a sentence parser typically used for?

 **It is used to parse sentences to derive their most likely syntax tree structures**

Suppose you have trained a logistic regression classifier and it outputs a new example ‘X’ with a prediction HO(X) = 0.2. This means what?

 **Our estimate for P(Y) = 0 for X**

What is pca.components\_ in SKlearn?

 **Set of all Eigen vectors for the projection space**

Which of the following is an example of a deterministic algorithm?

 **PCA**

A Pearson correlation between to variables is 0 but their values can still be related to each other?

 **True**

Imagine you are solving a classification problem with highly imbalanced class, the majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on the test data. Which of the following is true in such a case?

1. Accuracy matrix is not is good idea for imbalanced class problems
2. Accuracy matrix is a good idea for imbalanced class problems
3. Precision and recall matrix are good for imbalanced class problems
4. Precision and recall matrix are not good for imbalanced class problems

**Option 1 & 3 are correct**

Which of the following option is true for overall execution time for 5 fold cross validation with 10 different values of max\_depth?

 **More than 600 secs**

What would you do in PCA to get the same projection as SVM?

 **Transform data to zero mean.**

Which of the following value of K will have least leave-one-out cross validation accuracy?

 **1-NN**

Which of the following options can be used to get global minima K-means algorithm?

1. Try to run algorithm for different centroid initialization
2. Adjust number of iterations
3. Find out the optimal no. of clusters
4. **All of the above**

Imagine, you have a 28 \* 28 image and you run a 3 \* 3 convolution neural network on it with the input depth of 3 and output depth of 8.Note: Stride is 1 and you are using same padding.

A) 28 width, 28 height and 8 depthB) **13 width, 13 height and 8 depth**C) 28 width, 13 height and 8 depthD) 13 width, 28 height and 8 depth

A feature F1 can take certain values: A, B, C, D, E & F and represents grade of from a college.

Which of the following statement is true for the above case?

1. Feature F1 is an example of nominal variable
2. Feature F1 is an example of ordinal variable
3. It doesn’t belong to any of the above category
4. Both of these

**Case 2**

Assume that there is a blackbox algorithm which takes training data with multiple observations T1, T2, T3,………., Tn and a new observation Q1. The blackbox the nearest neighbour of Q1 say Ti and its corresponding class level Ci. Assume that this blackbox algorithm is same as 1- NN.

It is possible to construct a K-NN classification algorithm based on this blackbox alone where number of training observations is very large compared to K?

**True**

Assume that there is a blackbox algorithm which takes training data with multiple observations T1, T2, T3,………., Tn and a new observation Q1. The blackbox the nearest neighbour of Q1 say Ti and its corresponding class level Ci. Assume that this blackbox algorithm is same as 1- NN.

Instead of using 1-NN blackbox we want to use the J-NN algorithm for blackbox, where J>1?

Which of the following option is correct for finding K-NN using J-NN?

1. **J must be a proper factor of K.**
2. J must be greater than K.
3. Not possible