**MCQ**

1. As an input for calculating log loss for model evaluation of classification algorithm, we require
   1. Response variable values of test set(y\_test) and predicted probabilities from the algorithm on test data (output from predict\_proba(X\_test) )
   2. Response variable values of train set(y\_train) and predicted probabilities from the algorithm on test data(output from predict\_proba(X\_test) )
   3. Response variable values of test set(y\_test) and predictions from the algorithm on test data (output from predict(X\_test) )
   4. Response variable values of test set(y\_test) and predictions from the algorithm on train data(output from predict(X\_train) )

Ans: a

1. In which (unsupervised) clustering method is clustering of observations based on the nearness (neighbourhood) of the points
   1. K-Means
   2. Hierarchical
   3. DBSCAN
   4. K-NN

Ans: c

1. Min Max scaler transforms the data to the range
   1. (1,10)
   2. (-infinity, infinity)
   3. (0,infinity)
   4. (0,1)

Ans: d

1. There is an option in Python to calculate area under the curve of ROC Curve in case of Multi-Class problem.
   1. TRUE
   2. FALSE

Ans: a

1. What is different with the trees grown with random forest?
   1. Nothing different same as single decision tree
   2. The best split is decided by cost function
   3. The best split is decided only on a small number of randomly selected features
   4. The best split is decided only on a small number of randomly selected observations

Ans: c

1. The principal component analysis transformation ultimately gives us
   1. All the observations which are independent of each other
   2. All the features which are independent of each other
   3. All the observations which are dependent on each other
   4. All the features which are dependent on each other

Ans: b

1. Match the following ( 5 marks)

|  |  |
| --- | --- |
| ***Model*** | ***Important Hyper-parameters*** |
| 1. support vector machines - radial | a. k (n\_neighbors) |
| 2. decision trees | b. min\_samples\_split, min\_samples\_leaf, max\_depth |
| 3. k-nearest neighbours | c. learning\_rate , max\_depth, n\_estimators |
| 4. random forest | d. max\_features |
| 5. gradient boosting | e. C, gamma |

1. 1-a,2-b,3-c,4-d,5-e
2. 1-e,2-d,3-a,4-a,5-c
3. 1-e,2-b,3-a,4-d,5-c
4. 1-d,2-b,3-a,4-e,5-c

Ans: c

1. Which of the statements is true with parameters and hyper-parameters?
   1. Hyper-parameters are decided before the program is executed, whereas parameters are calculated from the data
   2. Parameters are decided before the program is executed, whereas hyper-parameters are calculated from the data
   3. Parameters and Hyper-parameters are same
   4. None of Above

Ans: a

1. Support Vector Machines in Python cannot be performed with Multi-Class classification problem
   1. TRUE
   2. FALSE

Ans: b

1. In the cost function J( ), of decision tree algorithm, which of the metric is not included?
   1. Area Under the Curve Score of ROC
   2. Gini’s Index
   3. Squared Error
   4. Log Loss

Ans: a

1. Concept of Simple random sampling with replacement is used in which of the ML algorithms?
   1. Support Vector Machines
   2. Linear Regression
   3. Bagging
   4. Stacking

Ans: c

1. The argument (random\_state=…) is specified
   1. So that every time the function is run, the output gets different
   2. So that every time the function is run, the output gets same
   3. Because it is compulsory argument
   4. None of these

Ans: b

1. Log loss in Python can be computed only for binary classification and not for multi-class classification
   1. TRUE
   2. FALSE

Ans: b

1. Which of the following is not evaluation metrics for classification algorithms?
   1. R2 score
   2. Area under ROC
   3. F1 Score
   4. Precision

Ans: a

1. The function in scikit-learn named GridSearchCV.fit( ) always finds the best parameter set
   1. based on maximum of the score mentioned in its argument option “scoring=”
   2. based on minimum of the score mentioned in its argument option “scoring=”
   3. depends upon ‘random\_state=’ option
   4. None of Above

Ans: a

1. K-Means clustering doesn’t start with randomization.
   1. TRUE
   2. FALSE

Ans: b