MACHINE LEARNING ASSIGNMENT - 4

1 In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:

C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?

D) Ridge Regularisation

3. Which of the following is not a kernel in Support Vector Machines?

C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

A) Logistic Regression

5. In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' will be? (1 kilogram = 2.205 pounds)

C) old coefficient of 'X' ÷ 2.205

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

B) increases

7. Which of the following is not an advantage of using random forest instead of decision trees?

C) Random Forests are easy to interpret

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. Which of the following are correct about Principal Components?

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables

9. Which of the following are applications of clustering?

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. Which of the following is(are) hyper parameters of a decision tree?

A) max_depth B) max_features D) min_samples_leaf

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. For eg: While analyzing salaries of all the people in India, salary of Mukesh Ambani or Ratan Tata might be an outlier. Inter Quartile Range (IQR): IQR = Q3-Q1 Where, Q1 = 25th % ile of the data Q2 = 50th % ile (a.k.a. median) Q3 = 75th % ile of the data. Upper bound = Q3 + 1.5*Q3 Lower Bound = Q1 – 1.5*Q1 Any data point lying above than upper bound and lower than lower bound is considered as an outlier.

12. What is the primary difference between bagging and boosting algorithms?

Bagging: Bagging is also known as bootstrap aggregating sits on top of the majority voting principle. The samples are bootstrapped each time when the model is trained. When the samples are chosen, they are used to train and validate the predictions. The samples are then replaced back into the training set. The samples are selected at random. This technique is known as bagging. To sum up, base classifiers such as decision trees are fitted on random subsets of the original training set. Subsequently, the individual predictions are aggregated (voting or averaging etc.). The final results are then used as predictions. It reduces the variance of a black box estimator. Due to this the chances of overfitting is ruled out. Boosting: The concept of Adaptive Boost revolves around correcting previous classifier mistakes. Each classifier gets trained on the sample set and learns to predict. The misclassification errors are then fed into the next classifier in the chain and are used to correct the mistakes until the final model predicts accurate results. When a weak-classifier misclassifies a training sample, the algorithm then uses these very samples to improve the performance of the ensemble.

13. What is adjusted R2 in linear regression. How is it calculated?

Adjusted R2 and R2 both represent that how well the model fits the data points. But adjusted R2 penalizes the model for using more features. In case we increase the number of features in training data the R2 will increase but adjusted R2 will only increase if the new feature adds value to our model. Due to this reason adjusted R2 is considered as a better evaluation metric than R2 . Adjusted R 2 is always less than or equal to R2 . The formula to calculate adjusted R2 is as follows: $Radj$ **2 = [1 − (1 −** R **2)(**n **− 1)** n **−** k **− 1] Where, n = number of data points in the dataset K = Number of features in the dataset excluding the constant term**

14. What is the difference between standardisation and normalisation?

In Normalization a dataset is scaled in such a way that all the data points lie between 0 and 1. Normalization is often called min-max scaling. Formula for Normalization is as follows: min max min new x x x x x − = − Whereas, In Standardization a dataset is scaled in such a way that the mean of data points becomes 0 and standard

deviation is 1. The transformed data may be positive as well as negative in standardization. The formula for standardization is as follows: i new x x x s − = Where, i x = ith data point x = sample mean s = sample standard deviation

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Cross validation is a technique to fit a model on data set. In cross validation the data set is divided into 'k' number of sets where 'k-1' sets are used for training and 1 set is used as validation set. And this is done for all the set one by one and the final score of model is taken as average score of all the 'k' number of fits. Advantage of using Cross validation is that, there is no need of separate validation data, cross validation reduces chances of overfitting and gives a more generic model. Cross validation has a disadvantage that it takes more time to fit the model over a large dataset and the model built is more complex than the basic model.