MACHINE LEARNING ASSIGNMENT - 4

**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?**

A) Lasso Regularisation B) PCA

C) Recursive feature elimination 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)

A) 2.205 × old coefficient of ‘X’

**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

**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

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

A data point that varies greatly from other observations is referred to as an outlier. An outlier may be caused by measurement uncertainty or by experimental error, the latter of which is often omitted from the data set. In statistical analyses, an outlier can cause serious problems.

The interquartile range (IQR), also known as the middle 50 percent or midspread, is a statistical dispersion measure that is equal to the gap between the 75th and 25th percentiles, or upper and lower quartiles, IQR = Q3 Q1.

The IQR is calculated by subtracting the first quartile from the third quartile.

It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers.

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

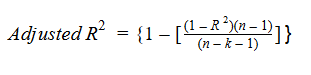
1. Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.
2. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.
3. In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.
4. Models are built independently in Bagging. New models are affected by a previously built model’s performance in Boosting.
5. In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. In Boosting, every new subset comprises the elements that were misclassified by previous models.
6. Bagging is usually applied where the classifier is unstable and has a high variance. Boosting is usually applied where the classifier is stable and simple and has high bias.

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

The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not.

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

Let’s have a look at the formula for adjusted R-squared to better understand its working.



N - represents the number of data points in our dataset

k - represents the number of independent variables, and

R - represents the R-squared values determined by the model

**14. What is the difference between standardisation and normalisation?**

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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

Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets.

Advantage - Reduces Overfitting  
Disadvantage - Increases Training Time