**MACHINE LEARNING**

**Answers for WORKSHEET – 1**

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 Regularization**

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

**C) Hyper plane**

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

**D) Support Vector Classifier**

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?

**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,C**

9. Which of the following are applications of clustering?

**A,B,C,D**

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

**A,B,D**

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

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Methods used for inter quartile range to detect outliers:

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

To calculate the inter quartile range by using the below formula:

Q3 – Q1

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

Bagging and Boosting decrease the variance of your single estimate as they combine several estimates from different models. So the result may be a model with **higher stability**.

If the problem is that the single model gets a very low performance, Bagging will rarely get a **better bias**. However, Boosting could generate a combined model with lower errors as it optimises the advantages and reduces pitfalls of the single model.

By contrast, if the difficulty of the single model is **over-fitting**, then Bagging is the best option. Boosting for its part doesn’t help to avoid over-fitting; in fact, this technique is faced with this problem itself. For this reason, Bagging is effective more often than Boosting.

* They are built independently for Bagging, Boosting tries to add new models that do well where previous models fail.
* Both are good at reducing variance and provide higher stability but only Boosting tries to reduce bias. On the other hand, Bagging may solve the over-fitting problem, while Boosting can increase it.

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

Adjusted R2 is the square value of R which can be placed near to certain value used for regression type of problem to calculate the accuracy of model.

Adjusted R2 is the More near value of R square value to 1 means the line is best.

This feature also tries to reduce the overfitting problem and can also determines to reduce the curse of dimension

**14.What is the difference between standardization and normalization ?**

**Normalization** usually means to scale a variable to have a values **between** 0 and 1, while **standardization** transforms data to have a mean of zero and a standard deviation of 1.

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

Cross validation is a technique for assessing how the statistical analysis generalizes to an independent dataset.

Advantage - prevents model from over-fitting the training data set.

Disadvantage - the cross validation process can become a lengthy one.