**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?**

*B) same as 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.*

*Quartiles divide a rank-ordered data set into four equal parts.The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.*

*Inter quartile range method to detect outliers:*

1. *Calculate the interquartile range for the data ,formula:IQR=Q3 – Q1*
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.*

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

* *Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to the different types.*
* *Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.*
* *In Bagging each model receives equal weight whereas in Boosting models are weighted according to their performance.*
* *In Bagging ,each model is built independently whereas in Boosting new models are influenced by performance of previously built models.*
* *In Bagging different training data subsets are randomly drawn with replacement from the entire training dataset. In Boosting every new subset contains the elements that were misclassified by previous models.*
* *Bagging tries to solve over-fitting problems while Boosting tries to reduce bias.*
* *If the classifier is unstable (high variance), then we should apply Bagging. If the classifier is stable and simple (high bias) then we should apply Boosting.*
* *Bagging is extended to Random forest model while Boosting is extended to Gradient boosting.*

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

* *The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.*
* *The adjusted R-squared increases only if the new term improves the model more than what would be expected by chance.*
* *It decreases when a predictor improves the model by less than what is expected by chance.*
* *It is always lower than the R-squared.*

*Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.*

*R2 = (1-R2)(N-1)/N-p-1*

*Where R2 = sample R sq ; P = Number of predictors ;N = Total number of sample.*

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

*Normalization usually means to scale a variable to have values between 0 and 1, while standardization transforms data to have a mean of 0 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 in which,instead of training our model on one training dataset, we train it on many subsets of the dataset.It is a great technique to deal with overfitting problems in various algorithms.*

*Advantage:*

* *Reduces Overfitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset.*

*Disadvantage:*

* *Increases Training Time & Needs Expensive Computation: Cross Validation drastically increases the training time since we have to train our model on multiple training sets instead of just one.Also, it is computationally very expensive in terms of processing power required.*