**MACHINE LEARNING WORKSHEET-4**

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

1. The value of correlation coefficient will always be:

A) between 0 and 1 B) greater than -1

C) between -1 and 1 D) between 0 and -1

ANS1. ‘C’

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

A) Lasso Regularisation B) PCA

C) Recursive feature elimination D) Ridge Regularisation

ANS2.’D’

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

A) linear B) Radial Basis Function

C) hyperplane D) polynomial

ANS3.’C’

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

boundaries?

A) Logistic Regression B) Naïve Bayes Classifier

C) Decision Tree Classifier D) Support Vector Classifier

ANS4.’A’

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’ B) same as old coefficient of ‘X’

C) old coefficient of ‘X’ ÷ 2.205 D) Cannot be determined

ANS5.’C’

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

the model?

A) remains same B) increases

C) decreases D) none of the above

ANS6.’B’

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

A) Random Forests reduce overfitting

B) Random Forests explains more variance in data then decision trees

C) Random Forests are easy to interpret

D) Random Forests provide a reliable feature importance estimate

ANS7.’C’

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?

A) Principal Components are calculated using supervised learning techniques

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

D) All of the above

ANS8.’D’

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

B) Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts.

C) Identifying spam or ham emails

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

levels.

ANS9.’A’ & ‘D’

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

A) max\_depth B) max\_features

C) n\_estimators D) min\_samples\_leaf

ANS10.’A’, ‘B’ & ‘D’

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

ANS11. An outlier is a data point which is considerably different from other points. Outliers can occur in any distribution by chance, but frequently either indicate a measurement error or a large-scale distribution in the population.

IQR is used to measure variability  by dividing a data set into 4 different quartiles. The data is sorted in ascending order and split into 4 equal parts.

Q1-25th percentile

Q2-50th percentile

Q3-75th percentile

IQR is the range between the first and the third quartiles namely Q1 and Q3: *IQR = Q3 – Q1*. The data points which fall below *Q1 – 1.5 IQR* or above *Q3 + 1.5 IQR* are outliers.

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

Ans12.In Bagging algorithms the final result is calculated by averaging the responses of the N learners (or majority vote).

However, Boosting algorithms , it assigns a second set of weights, this time for the N classifiers, in order to calculate the weighted average of their estimates.

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

Ans13.The adjusted R-squared is a transformed version of R-square that is used as an adjuster for predictors that do not show a significant value

Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model

14. What is the difference between standardisation and normalisation?

ANS14. Normalization means that the range of values are normalized to be from 0.0 to 1.0.

Standardization typically means that the range of values are standardized to measure how many standard deviations the value is from its mean

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

ANS15. Crossvalidation is a technique by which we use the data set sub-set to train our model and then compare the data set with the complementary subset.

Advantages

1. It reduces over fitting
2. It helps in hyperparameter tuning

Disadvantages

1. Increases Time required for Training the data
2. It requires expensive computation like processing power and speed