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:
A) between 0 and 1
B) greater than -1
C) between -1 and 1
D) between 0 and -1

Ans:- C

2. Which of the following cannot be used for dimensionality reduction?
A) Lasso Regularisation
B) PCA
C) Recursive feature elimination
D) Ridge Regularisation

Ans:-B

3. Which of the following is not a kernel in Support Vector Machines?
A) linear
B) Radial Basis Function
C) hyperplane
D) polynomial

Ans:-A

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

Ans:-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

Ans:-B

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

Ans:- 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

Ans:-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

Ans:-B

9. Which of the following are applications of clustering?

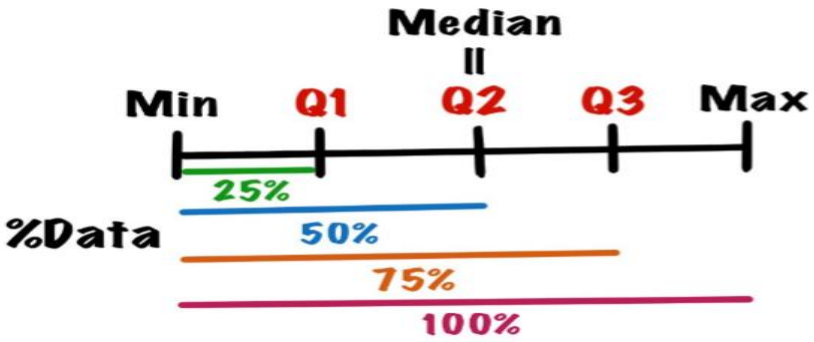
A) Identifying developed, developing and under-developed countries on the basis of factors like GDP,
poverty index, employment rate, population and living index
B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
C) Identifying spam or ham emails
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar
levels.

Ans:-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

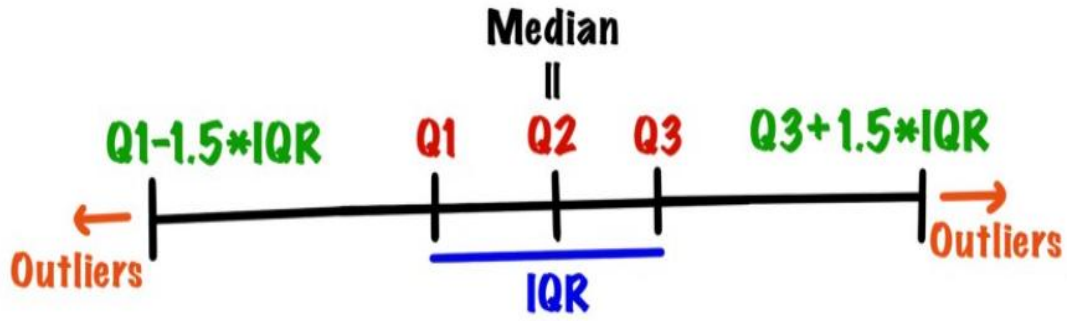
Ans:- A,B,D

MACHINE LEARNING
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.
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.
The interquartile range is a widely accepted method to find outliers in data. When using the interquartile range, or IQR, the full dataset is split into four equal segments, or quartiles. The distances between the quartiles is what is used to determine the IQR.
One common technique to detect outliers is using IQR (interquartile range). In specific, IQR is the middle 50% of data, which is Q3-Q1. Q1 is the first quartile, Q3 is the third quartile, and quartile divides an ordered dataset into 4 equal-sized groups.



So, here in figure we see that median value Q1,Q2,Q3 are Quartile 1,2,3. Inner quartile range is Q3 – Q1,as we discussed earlier.

The interquartile range method defines outliers as values larger than Q3 + 1.5 * IQR or the values smaller than Q1 – 1.5 * IQR.



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

- Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.

- Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

- In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.

- Models are built independently in Bagging. New models are affected by a previously built model's performance in Boosting.

- 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.

- 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?

Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.

R^2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects are included in the model. Adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might decrease if a specific effect does not improve the model.

The formula to calculate the adjusted R square of regression is below:

R^2 = {(1 / N) * Σ [(xi – x) * (Yi – y)] / (σx * σy)}^2

Where

- R^2= adjusted R square of the regression expression
- N= Number of observations in the regression equation
- Xi= Independent variable of the regression equation

- X= Mean of the independent variable of the regression equation
- Yi= Dependent variable of the regression equation
- Y=mean of the dependent variable of the regression equation
- σx = Standard deviation of the independent variable
- σy = Standard deviation of the dependent variable.

14. What is the difference between standardisation and normalisation?

NORMALIZATION:-

- Minimum value and maximum value of feature used for scaling
- It is used when feature are of different scale.
- Scale values are [0,1] or [-1,1].
- It is really affected by outliers.
- It is useful when we do not know about distribution

STANDARDISATION: -

- Mean and standard deviation is used for scaling
- It is used when we ensure the zero mean and unit standard deviation
- It is much less affected by the outliers
- It is useful when the the distribution is gaussian or normal
- It is often called the z_score normalization

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 generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Using cross-validation, there are high chances that we can detect over-fitting with ease

Advantage of cross validation:-

- More accurate estimate of out-of-sample accuracy.
- More “efficient” use of data as every observation is used for both training and testing.
- Cross-validation gives the idea about how the model will generalize to an unknown dataset.
- Cross-validation helps to determine a more accurate estimate of model prediction performance

Disadvantage of cross validation:-

- with cross-validation, we need to train the model on multiple training sets.
- Cross-validation is computationally very expensive as we need to train on multiple training sets.