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:

C) between -1 and 1

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

C) Recursive feature elimination

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

A) linear

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? (1 kilogram = 2.205 pounds)

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?

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

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

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.

Outliers are the observations that lies in abnormal distance from other data points or values in a random sample from a population.

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?

Bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It decreases the variance and helps to avoid overfitting. It is usually applied to decision tree methods.

Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model.

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

Adjusted R2 is a special form of R2, the coefficient of determination. It indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If more and more useless variables is added to a model, adjusted r-squared will decrease.

The formula is:

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

14. What is the difference between standardisation and normalisation?

Normalization is used to transform features to be on a similar scale. It is useful when there are no outliers as it cannot cope up with them. The scale generally ranges to [0,1 or [-1,1] sometimes.

X_new = (X - X_min)/(X_max - X_min)

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

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

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models.

Advantage: Reduces Overfitting

Disadvantage: Increases Training Time