ASSIGNMENT - 4

MACHINE LEARNING

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) between -1 and 1

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

A) Lasso Regularization

B) PCA

C) Recursive feature elimination

D) Ridge Regularization

Ans : C) Recursive feature elimination

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) linear

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. 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)

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: **C) old coefficient of 'X' ÷ 2.205**

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: C) decreases

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: A) Random Forests reduce overfitting

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 th above

Ans : B) Principal Components are calculated using unsupervised 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

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: 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.

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

C) n_estimators

D) min_samples_leaf

Ans: A) max_depth

B) max_features

D) min_samples_leaf

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.

Ans: An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value. When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, and Q3 called the first, second, and third quartiles are the values that separate the 4 equal parts. Q1 represents the 25th percentile of the data. Q2 represents the 50th percentile of the data. Q3 represents the 75th percentile of the data. If a dataset has 2n / 2n+1 data points, then Q1 = median of the dataset. Q2 = median of n smallest data points. Q3 = median of n highest data points. 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?

Ans:

Bagging	Boosting
Various training data subsets are randomly drawn with replacement from the whole training dataset.	Each new subset contains the components tha were misclassified by previous models.

Bagging attempts to tackle the over-fitting issue.	Boosting tries to reduce bias.
If the classifier is unstable (high variance), then we need to apply bagging.	If the classifier is steady and straightforwar (high bias), then we need to apply boosting.
Every model receives an equal weight.	Models are weighted by their performance.
Objective to decrease variance, not bias.	Objective to decrease bias, not variance.
It is the easiest way of connecting predictions that belong to the same type.	It is a way of connecting predictions that belor to the different types.
Every model is constructed independently.	New models are affected by the performance o the previously developed model.

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

Ans: Adjusted R squared is a modified version of Rsquared that has been adjusted for the number of predictors in the model. Adjusted Rsquared value can be calculated based on value of Rsquared. Every time you add an independent variable to a model, R squared increases, even if the independent variable is insignificant. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the Rsquared increases, even if the independent variable is insignificant. It never declines. The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

R^2 Sample R-Squared

N Total Sample Size

p Number of independent variable

14. What is the difference between standardization and normalisation?

Ans: Normalization is a part of data processing and cleansing techniques. The main goal of normalization is to make the data homogenous over all records and fields. It helps in creating a linkage between the entry data which in turn helps in cleaning and improving data quality. Whereas data standardization is the process of placing dissimilar features on the same scale. Standardized data in other words can be defined as rescaling the attributes in such a way that their mean is 0 and the standard deviation becomes 1.

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

Ans: Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. Cross validation is a technique for assessing how statistical analysis generalizes to an independent dataset. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on complementary subset of data.

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. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

Disadvantage - Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.