MACHINE LEARNING
ASSIGNMENT - 4

Submitted by: Anshul Dubey

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 -

Ans-between -1 and 1

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

A) Lasso Regularisation
B) PCA
C) Recursive feature elimination
D) Ridge Regularisation

Ans- Ridge Regularisation

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

A) linear
B) Radial Basis Function

C) hyperplane
D) polynomial

Ans- hyperplane

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-Logistic Regression

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-2.205 × old coefficient of 'X'

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-increases

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- Random Forests are easy to interpret

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-All of the above

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: Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

And

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: max_depth ,max_feature and 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:Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

Inter Quartile Range(IQR) is the range between third quartile Q3 and first quartile Q1. If a value lies in the IQR, it is not considered an outlier.

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

- 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?

Ans: It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes for adding independent variable that do not help in predicting the dependent variable. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

$$R2 \text{ adjusted} = 1 - (1-R2)(N-1)/(N-p-1)$$

R2 – Sample R square; p – number of predictors; N – total sample size

Every time we add an independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas Adjusted R-squared increases only when independent variable is significant and affects dependent variable.

14. What is the difference between standardisation and normalisation?

Ans: "Normalization" typically 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.

Ans: It is a method to ensure that important data is not left out during training. For example, If we train on 70% data there may be chances that some important data was present in the remaining 30% data. So, we try to eliminate this by using cross-validation; where we select some other data as training in the next iteration and the remaining for testing. Finally, a statistic measure like mean of all such iterations is taken. Types of cross-validation are:

- Exhaustive cross-validation
    - Leave-p-out cross-validation
    - Leave-one-out cross-validation
- Non-exhaustive cross-validation
    - k-fold cross-validation
    - Holdout method
    - Repeated random sub-sampling validation
- Nested cross-validation
    - k*l-fold cross-validation
    - k-fold cross-validation with validation and test set

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: Cross Validation is computationally very expensive in terms of time and processing power required.