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:

ANS. C) between -1 and 1

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

ANS. B) PCA

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

ANS. A) linear

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

ANS. A) 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)

Ans. A) 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?

ANS. B) increases

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

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?

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?

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?

Ans. A) max_depth

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 that lies an abnormal distance from other values in a random sample from a population.

In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

Method (IQR)

We can utilize the IQR strategy for recognizing exceptions to set up a "wall" beyond Q1 and Q3.

Any qualities that fall beyond this wall are viewed as anomalies. To assemble this wall we take 1.5 times the IQR and afterward deduct this worth from Q1 and increase the value of Q3.

This gives us the base and greatest wall posts that we contrast every perception with. Any perceptions that are more than 1.5 IQR underneath Q1 or more than 1.5 IQR above Q3 are viewed as exceptions.

This is the strategy that Minitab uses to distinguish exceptions of course.

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

BAGGING	BOOSTING
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. 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 R-squared increases, even if the independent variable is insignificant. It never declines.

Calculate linear regression R2?

$R\ 2 = 1 - \text{sum squared regression (SSR) total sum of squares (SST)}$, $= 1 - \sum (y\ i - y\ i\ \hat{}\ ) 2$ $\sum (y\ i - \bar{y}) 2$. The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared.

14. What is the difference between standardisation and normalisation?

Normalisation	Standardisation
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers
It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed
It is also known as Scaling Normalization	It is also known as Z-Score

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.

Advantage

- Cross-validation gives us an idea about how the model will perform on an unknown dataset.
- Cross-validation helps to determine a more accurate estimate of model prediction performance.

Disadvantage

- 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.