MACHINE LEARNING ASSIGNMENT - 4

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?
D) Ridge Regularisation

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

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

7. Which of the following is not an advantage of using random forest instead of decision trees?
D) Random Forests provide a reliable feature importance estimate

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?
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.
ALL OF THE ABOVE

10. Which of the following is(are) hyper parameters of a decision tree?
A) max_depth

D) min_samples_leaf

MACHINE LEARNING ASSIGNMENT -
4 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.
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, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.
Most commonly used method to detect outliers is visualization.

1.We use various visualization methods, like Box-plot, Histogram, Scatter Plot
2.Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier
3.Data points, three or more standard deviation away from mean are considered outlier
4.Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding
5.Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance..
Some of the most popular methods for outlier detection are:

- Z-Score or Extreme Value Analysis (parametric)
- Probabilistic and Statistical Modeling (parametric)
- Linear Regression Models (PCA, LMS)
- Proximity Based Models (non-parametric)
- Information Theory Models
- High Dimensional Outlier Detection Methods (high dimensional sparse data)

12. What is the primary difference between bagging and boosting algorithms?
Ans. In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.
While they are built independently for Bagging, Boosting tries to add new models that do well where previous models fail.
Only Boosting determines weights for the data to tip the scales in favor of the most difficult cases.
It is an equally weighted average for Bagging and a weighted average for Boosting, more weight to those with better performance on training data.
Only Boosting tries to reduce bias. On the other hand, Bagging may solve the over-fitting problem, while Boosting can increase it.

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

Ans. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not.
Every time you 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.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

Here,

- n represents the number of data points in our dataset
- k represents the number of independent variables, and
- R represents the R-squared values determined by the model.

So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease

14. What is the difference between standardisation and normalisation?
Ans. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

Ans. Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.
Advantages of cross-validation:

1. More accurate estimate of out-of-sample accuracy.
2. More "efficient" use of data as every observation is used for both training and testing.

There is a disadvantage because the cross validation process can become a lengthy one. It depends on the number of observations in the original sample and your chosen value of 'p.'