MACHINE LEARNING ASSIGNMENT - 4

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 (Increases if increased till hundreds then remains constant)

7. Which of the following is not an advantage of using random forest instead of decision trees?
C) Random Forests are easy to interpret

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.
Both B & C

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
C) Identifying spam or ham emails
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.
A, B & D

10. Which of the following is(are) hyper parameters of a decision tree?
A) max_depth
B) max_features
D) min_samples_leaf
A, B & D

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Answer: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Major causes for outliers in data are:

1) data entry or measurement errors
2) sampling problems and unusual conditions
3) natural variation

Interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data. Can also be called the midspread (middle 50% or fourth spread or H-spread). It is defined as the difference between the 75th and 25th percentiles of the data.

Q1 = First quartile = 25th percentile

Q3 = Third quartile = 75th percentile

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?

Answer:

Bagging	**Boosting**
Involves technique of parallel learning, ex: basically use multiple dtc parallelly	Involves series learning by clubbing multiple weaker models with different weights to boost decesion
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.	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 belong to the different types.

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

Answer: The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. It shows whether adding additional predictors improve a regression model or not.

1) Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model
2) Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

Adjusted R^2 is always less than or equal to R^2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0 indicates a model that has no predictive value.

It is calculated by dividing the residual mean square error by the total mean square error

Adj R^2 = {(1 / N) * Σ [(xi – x) * (Yi – y)] / (σx * σy)}^2

Adj R^2= adjusted R square of the regression equation

N= Number of observations in the regression equation

Xi= Independent variable of the regression equation

X= Mean of the independent variable of the regression equation

Yi= Dependent variable of the regression equation

Y= Mean of the dependent variable of the regression equation

σx = Standard deviation of the independent variable

σy = Standard deviation of the dependent variable.

Adj R^2 = 1 – { (1 – R^2) * (N - 1) / (N – p - 1) }

Adj R^2= adjusted R square of the regression equation

N= Number of observations in the regression equation

R^2= R square of the regression equation

p = Number of independent variable

14. What is the difference between standardisation and normalisation?

Answer:

Standardisation	Normalisation
Standardization is the subtraction of the mean and then dividing by its standard deviation	Normalization is the process of dividing of a vector by its length and it transforms your data into a range between 0 and 1
Changing the range of your data	Changing the shape of the distribution of data
Mean and standard deviation is used for scaling.	Minimum and maximum value of features are used for scaling
It is not bounded to a certain range.	Scales values between [0, 1] or [-1, 1].
It is much less affected by outliers.	It is really affected by outliers.

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

Answer: Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data. Use cross-validation to detect overfitting.

Advantage:

1) It gives your model the opportunity to train on multiple train-test splits. This gives you a better indication of how well your model will perform on unseen data.
2) Thus helps to detect overfitting and then we can reduce overfitting by parameter tuning for other techniques

Disadvantage:

1) The disadvantage of this method is that the training algorithm has to be rerun from scratch many times, which means much computation time and power.