Machine Learning Questions

Q1. The value of correlation coefficient will always be:

Ans. (C) between -1 and 1

Q2. Which of the following cannot be used for dimensionality reduction?

Ans. (D) Ridge Regularisation

Q3. Which of the following is not a kernel in Support Vector Machines?

Ans. (C) hyperplane

Q4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Ans. (D) Support Vector Classifier

Q5. 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. 2.205 × old coefficient of ‘X’

Q6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

Ans. increases

Q7. Which of the following is not an advantage of using random forest instead of decision trees?

Ans. Random Forests reduce overfitting

Q8. Which of the following are correct about Principal Components?

Ans. (A) Principal Components are calculated using supervised learning techniques

(B) Principal Components are calculated using unsupervised learning techniques

Q9. Which of the following are applications of clustering?

Ans. (C) Identifying spam or harm emails.

(D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Q10. Which of the following is(are) hyper parameters of a decision tree?

Ans. (A) max\_depth (B) max\_features (D) min\_samples\_leaf

Q11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Ans. **Outliers:** The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.  
However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.

**Inter Quartile Range (IQR):** 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.

* 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.

Q12. What is the primary difference between bagging and boosting algorithms?

Ans.

Q13. What is adjusted R2 in logistic regression. How is it calculated?

Ans.

Q14. What is the difference between standardisation and normalisation?

Ans. **Normalization diff. Standardization :** The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a z-score, and data points can be standardized.

Q15. 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.

The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

**Advantages of cross-validation:**

1. More accurate estimate of out-of-sample accuracy.
2. Reduces Overfitting

**Disadvantages of cross-validation:**

1. **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.