Machine Learning Set 4

Q1- The value of correlation coefficient will always be:
Answer- C) between -1 and 1

Q2- Which of the following cannot be used for dimensionality reduction
Answer-None of the above

Q3- Which of the following is not a kernel in Support Vector Machines?
Answer- C) hyperplane

Q4- Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
Answer- A) Logistic Regression

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?
Answer- C)old coefficient of 'X' ÷ 2.205

Q6- As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model
Answer- B) increases

Q7- Which of the following is not an advantage of using random forest instead of decision trees?
Answer-C) Random Forests are easy to interpret

Q8- Which of the following are correct about Principal Components?
Answer- B) & C)

Q9-Which of the following are applications of clustering?
Answer-A) & C)

Q10-Which of the following is(are) hyper parameters of a decision tree
Answer- A) ,B)&D)

Q11-What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection
Answer- An observation which differs from an overall pattern on a sample dataset is called an outlier.

Inter Quartile Range- 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.
IQR is the range between the first and the third quartiles namely Q1 and Q3:

***IQR = Q3 – Q1*.**

We can use the **IQR method** of identifying outliers to set up a "**fence**" outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers.
The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

Q12- What are the primary difference between bagging and boosting algorithms?
Answer-
1)

- **Bagging** is a method of merging the same type of predictions.
- **Boosting** is a method of merging different types of predictions.

2)

- **In Bagging,** each model receives an equal weight
- **In Boosting**, models are weighed based on their performance.

3)

- Models are built independently in **Bagging**
- New models are affected by a previously built model's performance in **Boosting**.

4)

- **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.

Q13- What is adjusted R2 in linear regression.
Answer- **Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression**

Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

$$\text{Adjusted R Squared} = 1 - [((1 - R2) * (n - 1)) / (n - k - 1)]$$

Where,
n = No of Observations/Total Sample Size
R2= R-Squared
K= No of Features

Q14-What is the difference between standardization and normalization?
Answer-

Normalization	Standardization
Scaling is done by the highest and the lowest values. **Formula**, **Xnew=X-Xmin/ Xmax-Xmin** The process of arranging the data in a database is known as Normalization.	Scaling is done by mean and standard deviation. **Formula**, **Xstand=x-mean(x) / Standard Deviation(x)** Data standardization is a process in which the data is restructured in a uniform format.
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

Q15- What is cross-validation? Describe one advantage and one disadvantage of using cross-validation

Answer- Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against over fitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

Advantage:

Reduces Overfitting: I n 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:

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.