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?

B) PCA

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?

B) Naïve Bayes Classifier

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?

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?

B) increases

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

B) Random Forests explains more variance in data then decision trees

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

D) All of the above

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

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?

A) max\_depth C) n\_estimators D) min\_samples\_leaf

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

Outliers are extreme values that fall a long way outside of the other observations. The process of identifying outliers has many names in data mining and machine learning such as outlier mining, outlier modeling and novelty detection and anomaly detection.

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.

A) Q1 represents the 25th percentile of the data.

B) Q2 represents the 50th percentile of the data.

C) Q3 represents the 75th percentile of the data.

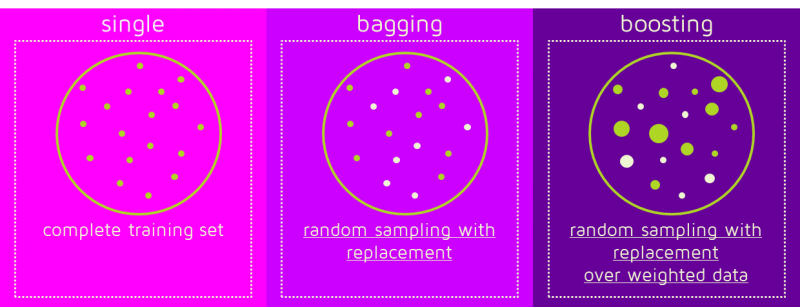
If a dataset has *2n / 2n+1* data point, then  
 Q1 = median of the dataset.  
 Q2 = median of n smallest data points.  
 Q3 = median of n highest data points.

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.

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

Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification

In the case of Bagging, any element has the same probability to appear in a new data set. However, for Boosting the observations are weighted and therefore some of them will take part in the new sets more often

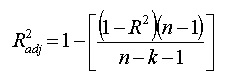


13. What is adjusted R2 in linear regression. How is it calculated?

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable.

R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/) to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase.  
Adjusted R2 will always be less than or equal to R2.

You only need R2 when working with [samples](https://www.statisticshowto.com/sample/). In other words, R2 isn’t necessary when you have data from an entire [population](https://www.statisticshowto.com/what-is-a-population/).

The formula is:  
  
where:

* N is the number of points in your data sample.
* K is the number of independent regressors, i.e. the number of [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/) in your model, excluding the [constant](https://www.calculushowto.com/constant-term-definition/).

14. What is the difference between standardization and normalization?

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.

Normalization equation

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.

Standardization equation

* Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
* Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Cross-validation is a technique for evaluating Machine Learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern.

Advantages of cross-validation:

* More accurate estimate of out-of-sample accuracy.
* More “efficient” use of data as every observation is used for both training and testing.

Disadvantage of cross-validation:

Cross validation is not a good choice, if you have multiple important confounders so you need to split independently of all of them, you end up with very small test and training set sizes for the surrogate models.