**MACHINE LEARNING**

Q1 to Q7 have only one correct answer. Choose the correct option to answer your question.

1. C) between -1 and 1
2. C) Recursive feature elimination
3. C) hyperplane
4. A) Logistic Regression
5. A) Yes
6. B) increases
7. Random Forests reduce overfitting

Q8 to Q10 are subjective answers type questions, Answers them in their own words briefly

1. B) Principal Components are calculated using unsupervised

learning techniques

C) Principal Components are linear combinations of Linear

Variables.

1. 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.

1. Max\_depth

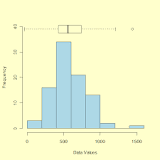
Max\_features

Q11 to Q15 are subjective answer type questions, Answer them briefly.

1. What are outliers? Explain the Inter Quartile Range (IQR)

method for outlier detection.

ANS: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.



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.

Example:  
Assume the data 6, 2, 1, 5, 4, 3, 50. If these values represent the number of chapatis eaten in lunch, then 50 is clearly an outlier.

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

ANS: Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

## Bagging

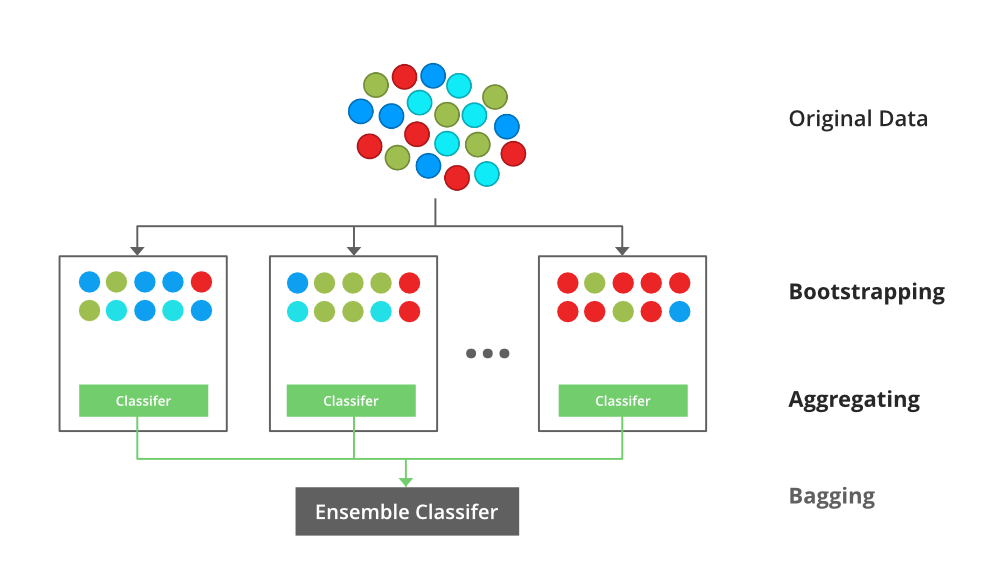
Bootstrap Aggregating, also known as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It decreases the [variance](https://www.geeksforgeeks.org/mathematics-mean-variance-and-standard-deviation/)and helps to avoid [overfitting](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/). It is usually applied to [decision tree methods](https://www.geeksforgeeks.org/decision-tree/). Bagging is a special case of the model averaging approach.

**Description of the Technique**

Suppose a set D of d tuples, at each iteration i, a training set Di of d tuples is selected via row sampling with a replacement method (i.e., there can be repetitive elements from different d tuples) from D (i.e., bootstrap). Then a classifier model Mi is learned for each training set D < i. Each classifier Mi returns its class prediction. The bagged classifier M\* counts the votes and assigns the class with the most votes to X (unknown sample).

Implementation Steps of Bagging

* Step 1: Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.
* Step 2: A base model is created on each of these subsets.
* Step 3: Each model is learned in parallel with each training set and independent of each other.
* Step 4: The final predictions are determined by combining the predictions from all the models.



## Boosting

Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

**Boosting Algorithms**

There are several boosting algorithms. The original ones, proposed by **Robert Schapire** and **Yoav Freund** were not adaptive and could not take full advantage of the weak learners. Schapire and Freund then developed [AdaBoost](https://www.geeksforgeeks.org/implementing-the-adaboost-algorithm-from-scratch/), an adaptive boosting algorithm that won the prestigious Gödel Prize. AdaBoost was the first really successful boosting algorithm developed for the purpose of binary classification. AdaBoost is short for Adaptive Boosting and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”.

1. Initialise the dataset and assign equal weight to each of the data point.
2. Provide this as input to the model and identify the wrongly classified data points.
3. Increase the weight of the wrongly classified data points and decrease the weights of correctly classified data points. And then normalize the weights of all data points.
4. if (got required results)  
     Goto step 5  
   else  
     Goto step 2
5. End

## https://media.geeksforgeeks.org/wp-content/uploads/20210707140911/Boosting.png

1. What is adjusted R2 in linear regression. How is it

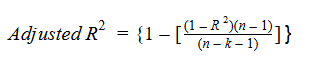
calculated?

ANS: Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

## Adjusted R-squared statistic

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

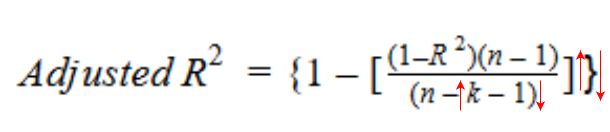
Let’s have a look at the formula for adjusted R-squared to better understand its working.



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.



you can calculate the adjusted R2 from R2 with a simple formula given

Adj r2 = 1-(1-R2)\*(n-1)/(n-p-1)

Where n is the sample size and p is the number of independent variables.

Adjusted R2 requires number of independent variables as well. That's why it will not be calculated using this function.

14. What is the difference between standardisation and normalisation?

ANS: In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

Normalisation

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:

[Normalization equation](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/Norm_eq.gif)

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

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:

[Standardization equation](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/Stand_eq.gif)

is the mean of the feature values andFeature scaling: Sigmais the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Normalization vs. standardization is an eternal question among machine learning newcomers.

**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.

# Use-case of Standardizer!

* In most of the Machine Learning models, it is been used & according to my & many other people's experiences, it outperforms MinMaxScaler(Normalization).
* Anywhere, where there is no need to scale features in the range 0 to 1.
* Since, it transforms the normal data distribution to standard normal distribution, which is the ideal & expected to have, most of the time it is the best to use in machine learning models.

# Use-case of Normalizer!

* Every situation where the range of features should be between 0 to 1. For example, in Images data, there we have color pixels range from 0 to 255(256 colors in total), here Normalizer is the best one to use.
* There can be multiple scenarios where this range is expected, there it is optimal to use MinMaxScaler.

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

ANS: Cross-validation is a technique that allows us to utilize our training data better for training and evaluating the model. For example, while using cross-validation, you effectively use complete data for training the model. Cross-validation also helps in finding the best hyperparameter for the model.

Advantages of Cross Validation  
  
1. Reduces Overfitting: In 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.  
  
Note: Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.  
  
2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

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.   
  
For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.  
  
2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.