**SOLUTIONS**

1.C)between -1 and 1

2.C) Recursive feature elimination

3.C)hyperplane

4.D) Support Vector Classifier

5. C) old coefficient of ‘X’ ÷ 2.205

6.A)remains same

7.C) Random Forests are easy to interpret

8.B) and C)

(B) Principal Components are calculated using unsupervised learning techniques and (C)Principal Components are linear combinations of Linear Variables)

9.A) and D)

(A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index and( D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.)

10.A) and D)

(A) max\_depth and (D) min\_samples\_leaf)

**Ans.11**.Outliers are anomalies,they are the extreme values in a dataset,that deviates significantly from the rest of the dataset,they may indicate variability in a measurement or execution error.In short outlier is a data point that lies outside the overall pattern in a distribution.The process of identifying outliers has many names in data mining and machine learning such as outlier mining,outlier modelling and novelty detection and ananomly detection.There are many ways to detect outliers such as –Z score,extreme value analysis(parametric),statistical modeling,IQR.

Let’s understand IQR method,

**Five number summary**,five number summary is a group of five numbers that help describe the shape and variation of a distribution.These five numbers are q2,the median of the set;q1 the median of lower half of the set;q3 the median of upper half of the set & maximum number that are outliers.

**A commonly used rule says that a data point is an outlier if it is more than 1.5.IQR above the third quartile or below the first quartlie.Said differently,low outliers are below Q1-1.5.**IQR and high outliers are above Q3+1.5.IQR.,Let’s take an example .here our 19 scores listed-5,7,10,15,19,21,21,22,22,23,23,23,23,23,24,24,24,24,25

Median-23(since 19 is odd median will be the middle score listed)

Q1(the first quartile)-19(the first quartile is the middle point of the 9 scores listed below the median)

Q3(the third quartile)-24(the third quartile is the middle point of the 9 scores above the median).

IQR=Q3-Q1=24-19 = 5

Let’s calculate 1.5.IQR below the first quartile and check for low outliers.

Q1-1.5.IQR=19-1.5(5)=19-7.5=11.5

Similarly for high outliers,

Q3+1.5.IQR=24-1.5(5)=24+7.5=31.5,there are 3 low outliers whereas no high outliers.In the similar way while plotting a box plot we can detect the outliers.

**Ans.12**.**Bagging and boosting are ensemble methods.the ensemble takes part in a bigger group of methods,called multiclassifiers**,where a set of hundreds or thousands of learners with a common objective are fused together to solve the problem. The main cause of error in learning are due to noise,bias and variance.Ensemble helps to minimize these factors.These methods are designed to improve the stability and accuracy of machine learning algorithms.Bagging and boosting get N learners by generating additional data in the training stage.N new training data sets are produced by random sampling with replacement from the original set.By sampling with replacement some observations may be repeated in each new training data set.In case of bagging,any element has the same probability to appear in new data set.However ,for boosting the observations are weighted and therefore some of them will take part in the new sets more often

**The main difference between the two methods is that while the training stage is parallel for bagging (i.e each model is build independently).Boosting builds the new learner in a sequential wayNow coming to classification stage,In boosting algorithms each classifier is trained on data,taking account the previous classifier’s success.After each training step,the weights are redistributed.Misclassified data increases it’s weights to emphasise the most difficult cases.**

**However,Boosting assigns a second set of weights,this time for the N classifiers,In order to take a weighted average of their estimates.**If the problem is that the single model gets a very low performance ,Bagging will rarely get a better bias .However boosting could generate a combined model with lower errors as it is optimises the advantages and reduces pitfalls of the single model.By contrast ,if the difficulty of the single model is over fitting,in fact this technique is faced with this problem itself.For this reason ,Bagging is effective more than Boosting.

To sum up,

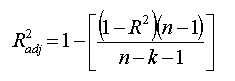
1.the main difference between bagging and boosting is that in bagging each model **is build independently,whereas boosting tries to add new models that do well where previous models fail**.(Bagging is a parallel technique ,boosting is a sequential technique)

2.Only **boosting determines weights for the data to tip the scales in favour of most difficult cases** unlike bagging where there is equally weighted average

3.**Boosting tries to reduce bias,while bagging solves over fitting problem.**

**Ans.13**.R squared shows how well terms(data points) fit a curve or line. It is also known as the coefficient of determination. **Adjusted R squared also indicates how well terms fit a curve or line,but adjusts for the number of terms in a model.**If we add more and more useless variables to a model,adjusted r-squared will decrease.**If we add more useful variables ,adjusted r-squared will increase.**Thereby adjusted r squared becomes important in logistic regressions because we have to classify some useful target variables(that is some categories) .Both r squared and the adjusted r squared gives us an idea of how many data points fall within the line of the regression.However there is one main difference between r squared and the adjusted r squared ,that is r squared assumes that every single variable explains the variation in the dependent variable,whereas the **adjusted R squared tells us the percentage of variation explained by only the independent variables that actually affect the dependent variable**.The adjusted r squared penalizes for adding independent variables (K in the equation) that do not fit the model,some of these variables will be significant but you can’t be sure that significance is just by chance.The adjusted r square will compensate for this by that penalizing for those extra variables.R squared increases with every predictor added to a model.As r squared always increases and never decreases,it can appear to be a better fit with the more terms we add to the model.This can be completely misleading.Similarly if our model has too many terms and many high order polynomials we can run into problem of over fitting the data.When we over fit the data ,a misleading r squared can lead to misleading projections.

**The formula is,**

[](https://www.statisticshowto.com/wp-content/uploads/2013/09/r-squared-adjusted.jpg),

Where,

\*N is the number of points in our dataset

\*K is the number of independent regressors (that is the number of variables in our model,excluding the constant)

\*R squared that is the coefficient of determination (to calculate the total variance,we would subtract the average actual value from each of the actual values,square the results and sum them.from there,divide the first sum of errors(explained variance)by the second sum (total variance) subtract the result from one,and we have r squared.

**Ans.14.Normalisation and standardization are scaling techniques.**

**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.

**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.Now the question comes which to use where so,

Normalization is good to use when you know that the distribution of your data does not follow a Guassian distribution.This can be useful in algorithms that do not assume any distribution of the data like k nearest neighbours and neural networks.Standardization,on the other hand can be helpful in cases where the data follows a Guassian distribution .However this does not have to be neccessarilly true.Also,unlike normailisation ,standardisation does not have a bounding range.So,even if we have outliers in our data,they will not be affected by standardization.

So to sum up,

1.We can conclude that **using normalisation in our dataset we generate smaller standard deviations,than using standardisation method,**it implies the data is more concentrated around the mean if we scale data using normalisation(min max scaler)

2. **Standardisation is more robust to outliers and in many cases it is preferable over normalization.**

**Ans.15**.**Cross validation is a technique for assessing how the statistical analysis generalizes to an independent dataset**.,Normally in any prediction problem ,our model works on a known dataset.We can also call it the training dataset.However,in real time ,our model will have to work on an unknown dataset.

Cross validation is a technique for assessing how the statistical analysis generalizes to an independent dataset.

We shall now dissect the definition and reproduce it in a simple manner.

Before testing out any model, would you not like to test it with an independent dataset? Normally, in any prediction problem, your model works on a known dataset. You also call it the training dataset. However, in real-time, your model will have to work on an unknown dataset.The testing is what we refer to as cross validation.Once our model passes this test,it is fit too work anywhere.So ,we can say that the purpose of cross validation is to assess how our prediction model performs with an unkown dataset.

Advantage-

**Reduces overfitting,in cross validation we split the dataset into multiple folds and train the algorithms on different folds.**This prevents our model from overfitting the training dataset.So,in this way ,the model attains the generalization which is a good sign of a robust algorithm.**Also cross validation helps in finding the optimal value of hyperparameters** to increase the efficiency of the algorithm.(hyperparameter tuning).

Disadvantage

**Increases training time**,cross validation drastically increases the training time.Earlier we had to train our model only on one training set,but with cross validation we have to train our model on multiple training sets.cross validation **also needs expensive computation**,cross validation is computationally very expensive **in terms of processing power required.**