MACHINE LEARNING

1) C) between -1 and 1
2) D) Ridge Regularisation
3) A) linear
4) C) Decision Tree Classifier
5) B) same as old coefficient of ‘X
6) C) decreases
7) B) Random Forests explains more variance in data then decision trees
8) B) Principal Components are calculated using unsupervised learning technique
9) A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
   B) Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts. C) Identifying spam or ham emails
10) A) max_depth B) max_features D) min_samples_leaf

11) An **outlier** is an observation that lies abnormally far away from other values in a dataset. Outliers can be problematic because they can affect the results of an analysis. The interquartile range, often abbreviated IQR, is the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a dataset. It measures the spread of the middle 50% of values.

One popular method is to declare an observation to be an outlier if it has a value 1.5 times greater than the IQR or 1.5 times less than the IQR.

Find the Lower and Upper Limits

The lower limit is calculated as:

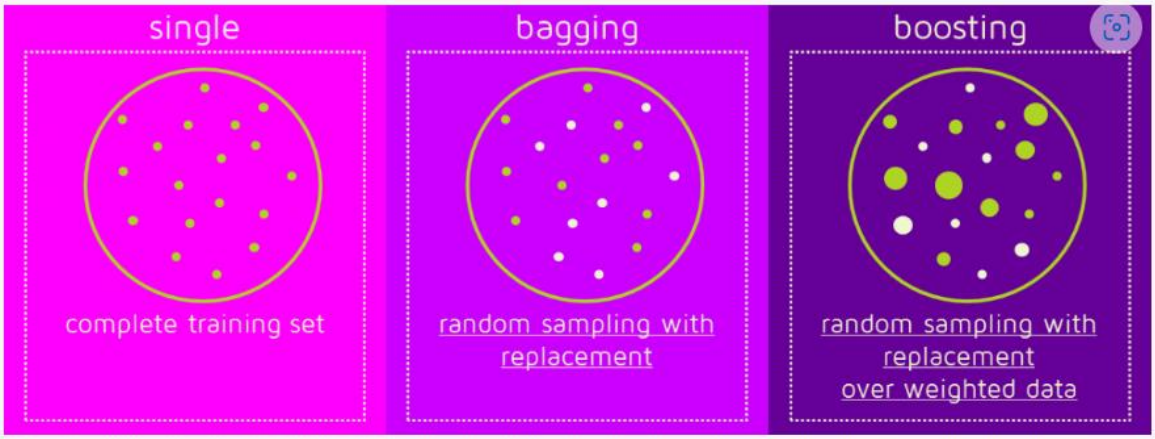
Lower limit = Q1 – 1.5*IQR

And the upper limited is calculated as:

Upper limit = Q3 + 1.5*IQR

12) 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 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:



These multiple sets are used to train the same learner algorithm and therefore different classifiers are produced.

13) The R-squared value is the proportion of the variance in the response variable that can be explained by the predictor variables in the model.

The value for R-squared can range from 0 to 1 where:

- A value of **0** indicates that the response variable cannot be explained by the predictor variables at all.
- A value of **1** indicates that the response variable can be perfectly explained by the predictor variables.

Although this metric is commonly used to assess how well a regression model fits a dataset. The **adjusted R-squared** is a modified version of R-squared that adjusts for the number of predictors in a regression model.

It is calculated as:

Adjusted $R^2 = 1 – [(1-R^2)*(n-1)/(n-k-1)]$

where:

- **R^2**: The R^2 of the model
- **n**: The number of observations
- **k**: The number of predictor variables

Because R-squared always increases as you add more predictors to a model, the adjusted R-squared can tell you how useful a model is, *adjusted for the number of predictors in a model*.

14) **Standardization** and **normalization** are two ways to rescale data.

Standardization rescales a dataset to have a mean of 0 and a standard deviation of 1. It uses the following formula to do so:

$$\mathbf{x_{new} = (x_i - x) / s}$$

where:

- $\mathbf{x_i}$: The i^{th} value in the dataset
- $\mathbf{x}$: The sample mean
- $\mathbf{s}$: The sample standard deviation

Normalization rescales a dataset so that each value falls between 0 and 1. It uses the following formula to do so:

$$\mathbf{x_{new} = (x_i - x_{min}) / (x_{max} - x_{min})}$$

where:

- $\mathbf{x_i}$: The i^{th} value in the dataset
- $\mathbf{x_{min}}$: The minimum value in the dataset
- $\mathbf{x_{max}}$: The maximum value in the dataset

15) Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:

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.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal

value of hyperparameters to increase the efficiency of the al

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.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.