Parth Makwana (DS0722)

Machine Learning

Q1 to Q7

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?
D) Ridge Regularisation

3. Which of the following is not a kernel in Support Vector Machines?
A) linear

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?
C) decreases

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

Q8 to Q10

8. Which of the following are correct about Principal Components?
B) Principal Components are calculated using unsupervised learning techniques
C) Principal Components are linear combinations of Linear Variables.

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
C) Identifying spam or ham emails

10. Which of the following is(are) hyper parameters of a decision tree?
A) max_depth
B) max_features
D) min_samples_leaf

Q10 to Q15

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.
Outlier
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. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Parth Makwana (DS0722)

Interquartile Range Definition

The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

Interquartile Range Formula

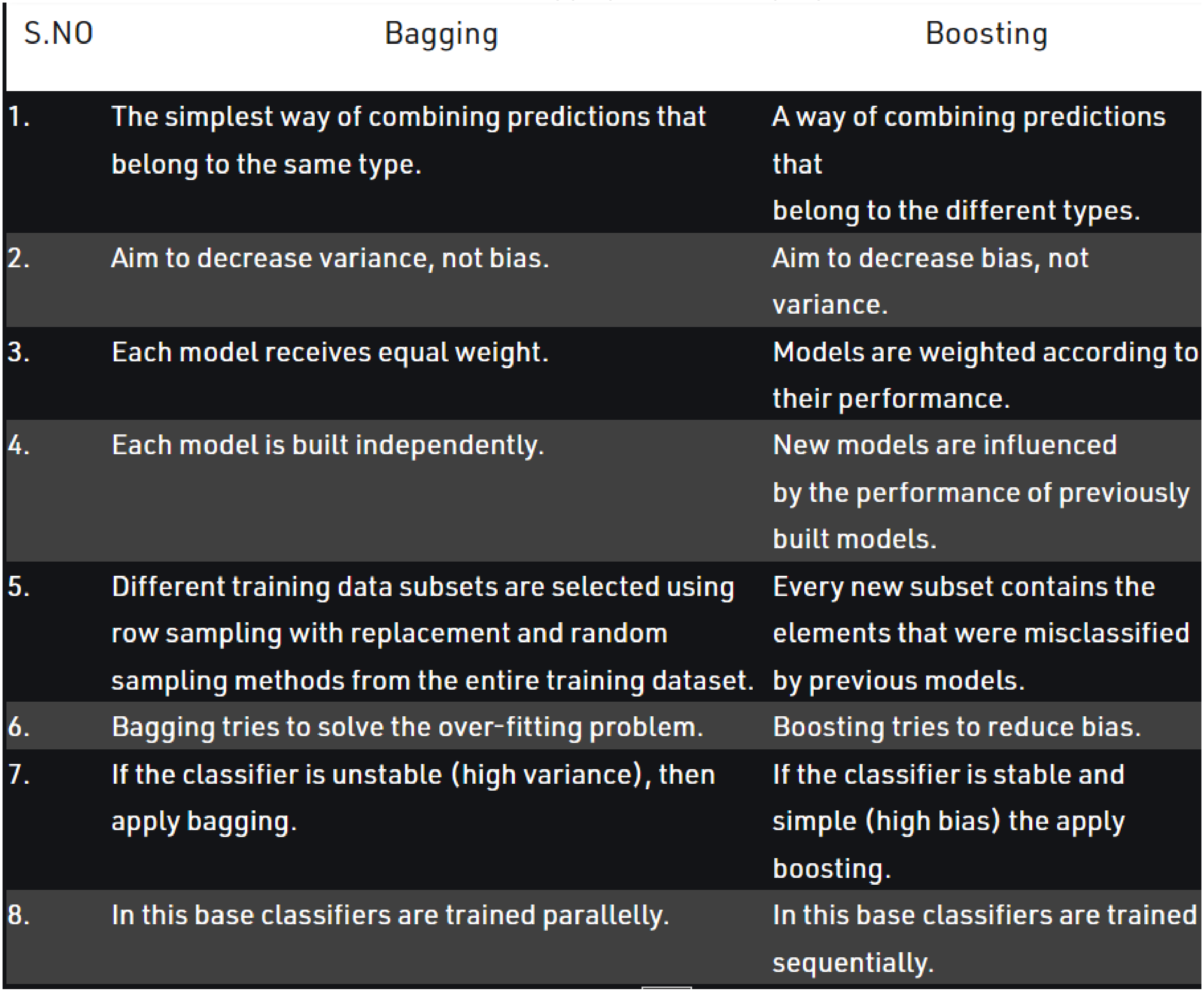
The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

Interquartile range = Upper Quartile – Lower Quartile = Q-3 – Q-1

where Q1 is the first quartile and Q3 is the third quartile of the series.

The below figure shows the occurrence of median and interquartile range for the data set.

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



S.NO	Bagging	Boosting
1.	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2.	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3.	Each model receives equal weight.	Models are weighted according to their performance.
4.	Each model is built independently.	New models are influenced by the performance of previously built models.
5.	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.
6.	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7.	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8.	In this base classifiers are trained parallelly.	In this base classifiers are trained sequentially.

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

The adjusted R-squared is a modified version of Rsquared that adjusts for the number of predictors in a regression model. When we fit linear regression models, we often calculate the R-squared value of 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.

R-squared will always increase when a new predictor variable is added to the regression model. It’s possible that a regression model with a large number of predictor variables has a high R-squared value, even if the model doesn’t fit the data well. So, to handle this there is an alternative to R-squared known as adjusted R-squared.

Formula:
Adjusted R2 = 1 – [(1-R2) * (n-1)/(n-k-1)]
where:

- R2: The R2 of the model
- n: The number of observations
- k: The number of predictor variables

14. What is the difference between standardisation and normalisation?

Normalization:	Standardization:
• Normalization or Min-Max Scaling is used to transform features to be on a similar scale. • Minimum and maximum value of features are used for scaling. • It is used when features are of different scales. • Scales values between [0, 1] or [-1, 1]. • It is really affected by outliers. • This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. • It is useful when we don’t know about the distribution. • It is an often called as Scaling Normalization • Formula: X_new = (X - X_min)/(X_max - X_min)	• Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. • Mean and standard deviation is used for scaling. • It is used when we want to ensure zero mean and unit standard deviation. • It is not bounded to a certain range. • It is much less affected by outliers. • It translates the data to the mean vector of original data to the origin and squishes or expands. • It is useful when the feature distribution is Normal or Gaussian. • Formula: X_new = (X - mean)/Std

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

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set. The three steps involved in cross-validation are as follows:

- Reserve some portion of sample data-set.
- Using the rest data-set train the model.
- Test the model using the reserve portion of the data-set.

Methods of Cross Validation:

- Validation Set Approach
- Leave-P-out cross-validation
- Leave one out cross-validation
- K-fold cross-validation

- Stratified k-fold cross-validation

The Advantages of CV are as follows:

- CV assists in realizing the optimal tuning of hyperparameters (or model settings) that increase the overall efficiency of the ML model,
- Training data is efficiently utilized as every observation is employed for both testing and training.
- More accurate estimate of out-of-sample accuracy.
- More “efficient” use of data as every observation is used for both training and testing.

The Disadvantages of CV are as follows:

- Increases Testing and Training Time: CV significantly increases the training time required for an ML model. This is due to the numerous test cycles to be implemented along with the test preparation and examining and analyzing of the results.
- Additional computation equates to additional resources required: CV is computationally expensive, requiring surplus processing power; add the first disadvantage of extra time, then this resource requirement will add further cost to an ML model project's budget.
- For the ideal conditions, it provides the optimum output. But for the inconsistent data, it may produce a drastic result. So, it is one of the big disadvantages of cross-validation, as there is no certainty of the type of data in machine learning.
- In predictive modeling, the data evolves over a period, due to which, it may face the differences between the training set and validation sets. Such as if we create a model for the prediction of stock market values, and the data is trained on the previous 5 years stock values, but the realistic future values for the next 5 years may drastically different, so it is difficult to expect the correct output for such situations