Homework Assignment – 3

Avisha Singh

Section A:

2. What metrics can you use to evaluate decision tree regression problem?

Ans- The metrics that can be used to evaluate decision tree regression problem -

• Mean absolute error (MAE):

- Average of absolute value of difference between actual and predicted values.

- Also referred to as loss function since the goal is to minimize this loss function.

Mean squared error (MSE):

- Considered as a loss function that needs to be minimized

- Heavily used in real-world ML applications

• Root mean squared error (RMSE):

- RMSE is basically the square root of MSE.

- Very popular loss function because of interpretive capability.

3. Why do we want a less complicated decision tree?

Ans- Tree-based methods are prone to overfitting.

• Decision trees, if allowed to grow complicated, they will generally overfit.

- It is better to simplify the tree to a smaller tree with fewer splits
    - lower model variance
    - better interpretation
    - with little added model bias

4. What are some of advantages and disadvantages of decision trees?

Ans- Advantages of decision trees-

- Easy to explain
- Analog to human decision making
- Graphically displayed
- Continuous or categorical variables

Disadvantages of decision trees-

- Lower predictive accuracy than other machine learning methods
- Model variance may be high

5. In your own word explain Bootstrap Aggregating (Bagging).

Ans- Bootstrap Aggregating (Bagging) is a type of ensemble technique in machine learning which helps to improve the accuracy of ML algorithms which is used for both classification and regression problems.

Working – We have a training dataset of particular size, in bagging new base learners (training sets) are generated. We will provide some random data samples to each of the base learner with the help of row sampling with replacement. Then all the base learners will get trained on those samples and will generate results, this step is called bootstrap.

The result which comes majority of the time will be considered as the actual result, this step is called aggregation.

6. True or False -

6.1. RMSE for Training Data predictions is 5.7 and RMSE for Testing Data predictions is 6.16. Then our model is overfitting.

Ans- True

6.2. Regression analysis can only be used for prediction.

Ans- False

6.3. Residuals are the difference between the observed values of y and the fitted values of y.

Ans- True

7. What is Machine learning? what is statistics? how do they differ? how are they the same? Give an example of a business question where a machine learning approach is most appropriate. Conversely give an example of where a statistical modeling approach is most appropriate.

Ans- Machine learning (ML) helps perform a specific task by detecting generalizable predictive patterns in data without being explicitly programmed to do so. ML allows you to perform inference and prediction. ML allows you to work with complicated data sets/big data analytics.

Statistics is the process of collecting, organizing, and interpreting data as well as drawing conclusions and making decisions.

Machine learning helps to detect generalizable predictive patterns whereas statistics is used to draw conclusions and make decisions from them.

Both machine learning and statistics share the same goal and objective that is learning from the data. Both of the approaches have a lot to learn from each other.

Machine learning is suitable to predict operational outcomes whereas statistical modeling can be used for analysis of micro seismic and seismic data.

8. What statistical parameters represent a normal distribution?

Ans- The statistical parameters which represent a normal distribution are mean which is responsible for symmetry of the graph and standard deviation which is the dispersion from the mean.

9. What could be some issues if the distribution of data is significantly different between test and training? and what might cause this difference?

Ans- The inference that we derive from the model can be flawed if our data is numerically significant and model can be overfitting model.

The difference can be caused due to design of train and test sample.

If the data is statistically different, then it is due to the inappropriate tests and sample sizes because when the size is increased the sensitivity also increases with that.

10. Explain the difference between classification and clustering.

Ans- Classification- It's a supervised learning technique.

- It's a process of classifying the data with help of class table.
- It's goal is to assign new input to a class.
- It works with labeled data.
- Known number of classes.

Clustering- It’s unsupervised learning technique.

- It is similar to classification but there are no predefined class table.
- Its goal is to find similarities within a given dataset.
- It works with unlabeled data.
- Unknown number of classes.