Assignment 05 Solutions

1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Ans: There are Five core tasks in the common ML workflow:

- Get Data: The first step in the Machine Learning process is getting data.

- Cleaning, Preparing & Manipulating Data: Real-world data often has unorganized, missing, or noisy elements.

- Train Model: This step is where the magic happens!

- Testing Model.

- Improving model.

- Data preprocessing involves transforming raw data to well-formed data sets so that data mining analytics can be applied.

- Preprocessing involves both data validation and data imputation

- The Goal of Data Validation is to assess whether the data in question is both complete and accurate.

- The Goal of Data Imputation is to correct errors and input missing values, Either Manually or Automatically through business process automation (BPA) programming.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two ?

Ans: The Data Type Is Broadly Classified Into:

1.Quantitative

2.Qualitative

- **Quantitative Data Type:** This Type Of Data Type Consists Of Numerical Values. Anything Which Is Measured By Numbers. E.G., Profit, Quantity Sold, Height, Weight, Temperature, Etc. This Is Again Of Two Types

  1. **Discrete Data Type:** – The Numeric Data Which Have Discrete Values Or Whole Numbers. This Type Of Variable Value If Expressed In Decimal Format Will Have No Proper Meaning. Their Values Can Be Counted. E.G.: – No. Of Cars You Have, No. Of Marbles In Containers, Students In A Class, Etc.

  2. **Continuous Data Type:** – The Numerical Measures Which Can Take The Value Within A Certain Range. This Type Of Variable Value If Expressed In Decimal Format Has True Meaning. Their Values Can Not Be Counted But Measured. The Value Can Be Infinite E.G.: Height, Weight, Time, Area, Distance, Measurement Of Rainfall, Etc.

- **Qualitative Data Type:** These Are The Data Types That Cannot Be Expressed In Numbers. This Describes Categories Or Groups And Is Hence Known As The Categorical Data Type. This Can Be Divided Into:-

  1. **Structured Data:** This Type Of Data Is Either Number Or Words. This Can Take Numerical Values But Mathematical Operations Cannot Be Performed On It. This Type Of Data Is Expressed In Tabular Format. E.G.) Sunny=1, Cloudy=2, Windy=3 Or Binary Form Data Like 0 Or1, Good Or Bad, Etc.

  2. **Unstructured Data:** This Type Of Data Does Not Have The Proper Format And Therefore Known As Unstructured Data.This Comprises Textual Data, Sounds, Images, Videos, Etc.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types ?

Ans: The following is a basic data collection that includes some sample records.

1. **Determine What Information You Want to Collect:** The first thing you need to do is choose what details you want to collect. You'll need to decide what topics the information will cover, who you want to collect it from and how much data you need. Your goals — what you hope to accomplish using your data — will determine your answers to these questions. As an example, you may decide to collect data about which type of articles are most popular on your website among visitors who are between the ages of 18 and 34. You might also choose to gather information about the average age of all of the customers who bought a product from your company within the last month.

2. **Set a Timeframe for Data Collection:** Next, you can start formulating your plan for how you'll collect your data. In the early stages of your planning process, you should establish a timeframe for your data collection. You may want to gather some types of data continuously. When it comes to transactional data and website visitor data, for example, you may want to set up a method for tracking that data over the long term. If you're tracking data for a specific campaign, however, you'll track it over a defined period. In these instances, you'll have a schedule for when you'll start and end your data collection.

3. **Determine Your Data Collection Method:** At this step, you will choose the data collection method that will make up the core of your data-gathering strategy. To select the right collection method, you'll need to consider the type of information you want to collect, the timeframe over which you'll obtain it and the other aspects you determined.

4. **Collect the Data:** Once you have finalized your plan, you can implement your data collection strategy and start collecting data. You can store and organize your data in your DMP. Be sure to stick to your plan and check on its progress regularly. It may be useful to create a schedule for when you will check in with how your data collection is proceeding, especially if you are collecting data continuously. You may want to make updates to your plan as conditions change and you get new information.

5. **Analyze the Data and Implement Your Findings:** Once you've collected all of your data, it's time to analyze it and organize your findings. The analysis phase is crucial because it turns raw data into valuable insights that you can use to enhance your marketing strategies, products and business decisions. You can use the analytics tools built into our DMP to help with this step. Once you've uncovered the patterns and insights in your data, you can implement the findings to improve your business.

4. What are the various causes of machine learning data issues? What are the ramifications ?

Ans: Noisy data, dirty data, and incomplete data are the quintessential enemies of ideal Machine Learning. The solution to this conundrum is to take the time to evaluate and scope data with meticulous data governance, data integration, and data exploration until you get clear data. Ramifications or major issues in machine learning are:

1. Five practical issues in machine learning and the business implications Data quality.

2. Machine learning systems rely on data.

3. The complexity and quality trade-off.

4. Sampling bias in data.

5. Changing expectations and concept drift.

6. Monitoring and maintenance.

5. Demonstrate various approaches to categorical data exploration with appropriate examples?

Ans: Various approaches to categorical data exploration are:

1. **Unique value count:** One of the first things which can be useful during data exploration is to see how many unique values

are there in categorical columns. 2. **Frequency Count:** Frequency count is finding how frequent individual values occur in column. 3. **Variance:** Variance gives a good indication how the values are spread. 4. **Pareto Analysis:** Pareto analysis is a creative way of focusing on what is important. Pareto 80–20 rule can be effectively used in data exploration. 5. **Histogram:** Histogram are one of the data scientists favourite data exploration techniques. It gives information on the range of values in which most of the values fall. It also gives information on whether there is any skew in data. 6. **Correlation Heat-map between all numeric columns:** The term correlation refers to a mutual relationship or association between two things. 7. **Pearson Correlation and Trend between two numeric columns:** Once you have visualised correlation heat-map , the next step is to see the correlation trend between two specific numeric columns. 8. **Outlier overview:** Finding something unusual in data is called Outlier detection (also known as anomaly detection). These outliers represent something unusual, rare , anomaly or something exceptional.

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Ans: Even in a Well-Designed & Controlled study, Missing data occurs in almost all research. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions.

- Real-world data collection has its own set of problems, It is often very messy which includes missing data, presence of outliers, unstructured manner, etc.

- Before looking for any insights from the data, we have to first perform preprocessing tasks which then only allow us to use that data for further observation and train our machine learning model.

- Missing value in a dataset is a very common phenomenon in the reality.

- Missing value correction is required to reduce bias and to produce powerful suitable models.

- Most of the algorithms can't handle missing data, thus you need to act in some way to simply not let your code crash. So, let's begin with the methods to solve the problem.

- Methods for dealing with missing values. The popular methods which are used by the machine learning community to handle the missing value for categorical variables in the dataset are as follows: Delete the observations: If there is a large number of observations in the dataset, where all the classes to be predicted are sufficiently represented in the training data, then try deleting the missing value observations, which would not bring significant change in your feed to your model.

7. Describe the various methods for dealing with missing data values in depth ?

Ans: The Various Methods for dealing with missing data values are:

1. **Delete the observations:** If there is a large number of observations in the dataset, where all the classes to be predicted are sufficiently represented in the training data, then try deleting the missing value observations, which would not bring significant change in your feed to your model. For Example Implement this method in a given dataset, we can delete the entire row which contains missing values.

2. **Replace missing values with the most frequent value:** You can always impute them based on Mode in the case of categorical variables, just make sure you don't have highly skewed class distributions.

3. **Develop a model to predict missing values:** One smart way of doing this could be training a classifier over your columns with missing values as a dependent variable against other features of your data set and trying to impute based on the newly trained classifier.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words ?

Ans: The Various Data Pre-Processing Techniques are:

- **Data Cleaning:** The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **Missing Data:** This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

- **Ignore the tuples:** This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

- **Fill the Missing values:** There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **Noisy Data:** Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

- **Binning Method:** This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

- **Regression:** Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

- **Clustering:** This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

- **Data Reduction:** Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. The various steps to data reduction are:

- **Data Cube Aggregation:** Aggregation operation is applied to data for the construction of the data cube.

- **Attribute Subset Selection:** The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

- **Numerosity Reduction:** This enable to store the model of data instead of whole data, for example: Regression Models.

- **Dimensionality Reduction:** This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction.

The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension.

9.Make brief notes on of the following ?

1. What is the IQR? What criteria are used to assess it?

2. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

Ans: The following is the brief notes on the following topics:

- **What is the IQR? What criteria are used to assess it?**

    - Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.

    - Q3 is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3.

    - The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR.

- **Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?**

    - minimum is the minimum value in the dataset

    - maximum is the maximum value in the dataset.

    - So the difference between the two tells us about the range of dataset.

    - The median is the median (or centre point), also called second quartile, of the data (resulting from the fact that the data is ordered).

    - Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.

    - Q3 is the third quartile of the data, i.e., to say 75%

    - When the data is left skewed, lower whisker will be longer than upper whisker.

    - To detect the outliers this method is used, we define a new range, let's call it decision range, and any data point lying outside this range is considered as outlier and is accordingly dealt with. The range is as given below:

Lower Bound: (Q1 - 1.5 * IQR)Upper Bound: (Q3 + 1.5 * IQR)

- The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR.

In [2]:

Out[2]:

10. Make brief notes on any two of the following ?

1. Data collected at regular intervals

2. The gap between the quartiles

3. Use a cross-tab

Ans: The following are the breif notes about:

- **Data collected at regular intervals:**

    - Interval data is one of the two types of discrete data.
    - An example of interval data is the data collected on a thermometer—its gradation or markings are equidistant.
    - Unlike ordinal data, interval data always take numerical values where the distance between two points on the scale is standardised and equal.

- **The gap between the quartiles:**

    - Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.
    - Q3 is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3.
    - The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR.

11. Make a comparison between ?

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

Ans: The following are the breif notes about:

- **The average and median:**

  - The mean (informally, the "average") is found by adding all of the numbers together and dividing by the number of items in the set: 10 + 10 + 20 + 40 + 70 / 5 = 30. The median is found by ordering the set from lowest to highest and finding the exact middle. The median is just the middle number: 20

- **Histogram and barplot:**

  - Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.

In [ ]: