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

Answer : There are five core tasks in the common ML workflow:

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

Clean, Prepare & Manipulate Data. Real-world data often has unorganized, missing, or noisy elements.

Train Model. This step is where the magic happens!

Test Model.

Improve.

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.

Answer : The Data Type Is Broadly Classified Into

1. Quantitative

2. Qualitative 1.

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

A.) 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.

B.) 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.

2. 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:-

A. 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.

B. 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.

Answer :

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

Answer : Data can also be noisy, filled with unwanted information that can mislead a machine learning model into making incorrect predictions.

The data used to train the model comes with its own biases. However, when we know the data is biased, there are ways to debias or to reduce the weighting given to that data.

Drift can occur when new data is introduced to the model. This is called data drift. It can also occur when our interpretation of the data changes.

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

Answer : 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?

Answer : If the missing values are not handled properly then we may end up drawing an inaccurate inference about the data. Due to improper handling, the result obtained will differ from ones where the missing values are present.

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.

This article was published as a part of the Data Science Blogathon.

Introduction

"Data is the fuel for Machine Learning algorithms".

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. In this blog, you will see how to handle missing values for categorical variables while we are performing data preprocessing. 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

Example 1, Let's have a **dummy dataset** in which there are three independent features(predictors) and one dependent feature(response).

Feature-1	*Feature-2*	*Feature-3*	*Output*
Male	23	24	Yes
– – – –	24	25	No
Female	25	26	Yes
Male	26	27	Yes

Here, We have a missing value in row-2 for Feature-1.

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:

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,1, Implement this method in a given dataset, we can delete the entire row which contains missing values(delete row-2).

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.

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

Answer :

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,1, Implement this method in a given dataset, we can delete the entire row which contains missing values(delete row-2).

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.

Answer :

1. 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.

- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:
  1. **Ignore the tuples:**
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

  2. **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.

(b). 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 :

1. **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.

2. **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).

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

3. 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:

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

2. **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.

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

4. **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.i. What is the IQR? What criteria are used to assess it?

ii. 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?

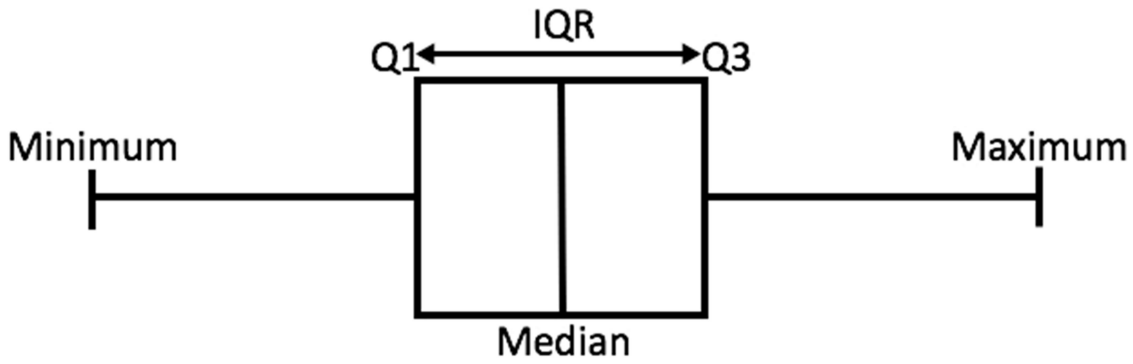
Answer : **i. What is the IQR? What criteria are used to assess it?**

- *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**.

ii. 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,
- and *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% of the data lies between *minimum* and *Q3*.

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**.

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

1. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

Answer : **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**.

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.