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

Machine Learning involves various steps listed below:-

* Data Collection
* Feature Engineering
* Feature Selection
* Model Building
* Model Deployment
* Improving & Model Retraining

Data pre-processing means data cleaning. The key tasks being employed by them are vectorization, normalisation and handling of null values. Vectorization is converting data in numerical format which can be easily understood by our machine. Normalisation involves converting all the feature values to be on a similar scale.

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

Quantitative data is anything that can be counted or measured; it refers to numerical data. Qualitative data is descriptive, referring to things that can be observed but not measured—such as colors or emotions.

|  |  |
| --- | --- |
| **Quantitative Data** | **Qualitative Data** |
| Countable or measurable, relating to numbers. | Descriptive, relating to words and language. |
| Tells us how many, how much or how often. | Describe certain attributes and helps us to understand the “why” and how behind certain behaviours. |
| Fixed and universal | Dynamic and subjective, open to interpretation |
| Gathered by measuring and counting things | Gathered through observations and interviews |
| Analysed using statistical analysis | Analysed by grouping data into meaningful categories |

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

Consider a job application involving following fields

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Age | Gender | Highest Qualification  (1-HSC, 2-BACHLEORS  3-MASTERS) | Percentage Obtained |
| Shantanu Ghosh | 36 | M | 3 | 77.5 |
| Vaibhavi Bisht | 32 | F | 2 | 82.3 |

Here the age column is discrete numerical data whereas percentage obtained is the continuous one. Gender is representing the categorical types whereas the highest qualification is of ordinal type.

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

The data issue being faced by a machine learning algorithm are lack of quality dataset, having too much noise in the dataset and data leakage.

Lack of quality dataset can be handled by fixing data in the source system, fixing the source system to correct data issues. Applying precision identity can also solve the problem.

Having too much noise in data can be solved by simply collecting more data.

Data leakage can be a problem which can lead to overfitting of our results. Usually in such a scenario, we often uses the methods of train test split.

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

The various approaches to categorical data exploration involves

* Loading the datasets.
* Gather info such as no. of rows, columns, data types and memory usage etc. by using .info().
* Checking for null values.
* Looking at the frequency distribution through visualisation using barplot, histogram, piechart etc.
* Converting categorical values into numerical through various techniques etc.:-
  + Replacing values
  + Encoding Labels
  + One hot encoding
  + Binary encoding
  + Backward difference encoding
  + Miscellaneous features

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

Learning activity will be flawed or the model will not be able to learn properly in case certain variables have missing values. The steps for treating them are discussed in next answer.

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

Null values can be treated in a number of ways mentioned below:-

1) Trimming/Removing the missing value rows.

2) Impute missing values with mean/median.

3) Impute for categorical variable can also be done. In this, whatever may be the categorical feature(string or number) , the missing value can be replaced with the highest occurring value. If the missing value is too much, it can be replaced with a completely new feature value.

4) Other Imputation Methods include forward filling the data having longitudinal behaviour. In time series dataset, we can use interpolation for filling the missing values.

5) Using algorithms that support missing values – KNN , Naïve bayes are used for making a prediction of missing values.

6) Prediction of missing values – Regression and classification model can also be used for making a prediction for missing values.

7) Imputation using Deep Learning Library – Datawig is a library that learns from ML model using deep neural networks to impute missing values in the dataset. This method works very well with categorical, continuous and non numerical features.

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

The various data pre-processing technique are :-

Data pre-processing means data cleaning. The key tasks being employed by them are vectorization, normalisation and handling of null values. Vectorization is converting data in numerical format which can be easily understood by our machine. Normalisation involves converting all the feature values to be on a similar scale.

Feature selection is selecting those features which are relevant to our dataset and excluding others while dimensionality reduction transforms the whole dataset into a lower dimension. Hence, whole of the information is retained.

**9.**

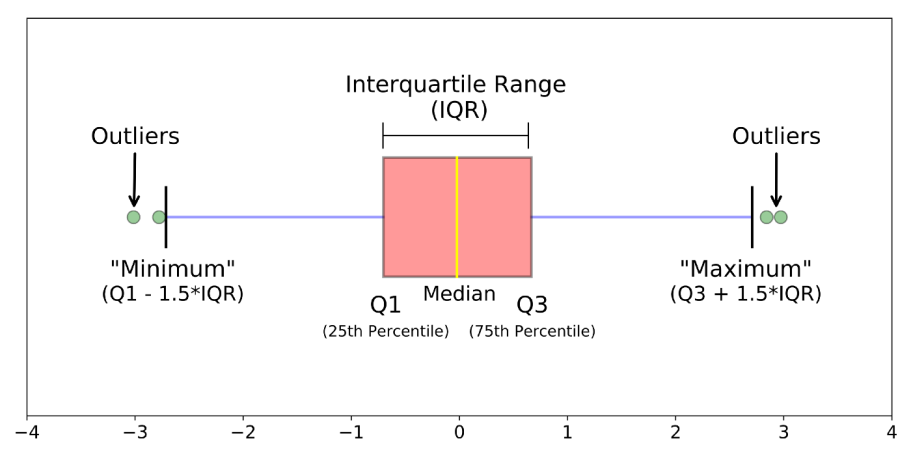
**i. What is the IQR? What criteria are used to assess it?**

IQR is interquartile range. It is basically represented by the difference between the upper quartile median and lower quartile median. It is a measure of statistical dispersion, which is the spreas of the datra. The steps involved in calculating it are as follows:-

* Order the data from descending to ascending.
* Find the median.
* Calculate the median of upper half data(Q3) and lower half data(Q1).
* IQR is the difference of upper and lower median.

IQR = Q3-Q1

**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?**



Median(Q2)=middle value in a dataset

First Quartile(Q1)=Middle value between the median and the smallest number.

Third Quartile(Q3)= Middle value between the median and the highest number.

IQR(InterQuartile Range) = Q3 – Q1

Whiskers – Whiskers are the line which shows the range of values from minimum to maximum.

The minimum point is given by Q2-(1.5\*IQR) and maximum by Q2+(1.5\*IQR)

If the lower whisker exceeds the upper whisker in length, then the median is closer towards the maximum point which means distribution of the dataset is said to be negatively skewed.

**10.** **Make brief notes on any two of the following:**

1. **Data collected at regular intervals**

This type of data collection involves analysing a sequence of data points collected over an equal interval of time. It shows how variable changes over a given interval of time. A large amount of data is required to be collected to ensure consistency and reliability. An extensive dataset ensures that we have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns are not outliers and can account for seasonal variance.

1. **The gap between the quartiles**

Quartiles divide the data points evenly. 25% of data points are present in each of the four quartiles. Median divides the set of data points into half. IQR is 50% of the data points between the third quartile and the first one.

3. Use a cross-tab

**11. Make a comparison between:**

1. **Data with nominal and ordinal values**

Nominal values is a naming scale, where variables are named or labelled with no specific order e.g. data with religious affiliation, sex, present city etc.

Ordinal values have all its variables in specific order beyond just naming them e.g. movie rating, military ranks etc.

1. **Histogram and box plot**

Histogram represents the frequency distribution of data along given class intervals. These class intervals are also known as bins. Histograms are commonly used in [statistics](https://www.investopedia.com/terms/s/statistics.asp) to demonstrate how many of a certain type of variable occurs within a specific range. For example, a census focused on the [demography](https://www.investopedia.com/terms/d/demographics.asp) of a country may use a histogram to show how many people are between the ages of 0 - 10, 11 - 20, 21 - 30, 31 - 40, 41 - 50, etc.

A box plot is used in explanatory data analysis. It is also known as box and whisker plot. It is a way of displaying the data through their quartiles.

1. **The average and median**

Average are arithmetic mean. It is given by the sum of all the values of the term divided by the total number of terms.

Median is the middle value in the dataset. If the dataset is

Odd, the median is

(Nth term/2)+((N+1)th term/2)

Otherwise,

(Nth term/2)