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

Answer: Following are the key tasks that machine learning entails:

1. Data collection
2. EDA
3. Data transformation and feature selection
4. Model building based on the use case and select the best model with best performance metrics and proceed for deployment.

Pre-processing of data simply means making some appropriate changes in the data like some small standard transformation so that our overall new transformed data helps our model to analyse some pattern going forward and make correct predictions.

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

Answer: Quantitative data is the information about quantities and so thereby about numbers, they’re collected by measuring things. Qualitative data is more descriptive and includes only those events which can be observed only and not measured, are collected by participant observation and information.

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:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Activity Date | Total Steps | Total Distance | Tracker Distance | Calories |
| 25/3/2016 | 11004 | 7.11 | 7.11 | 1819 |
| 26/3/2016 | 17609 | 11.55 | 11.55 | 2154 |
| 27/3/2016 | 12736 | 8.53 | 8.53 | 1944 |
| 28/3/2016 | 13231 | 8.93 | 8.93 | 1932 |
| 29/3/2016 | 12041 | 7.85 | 7.85 | 1996 |

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

Answer: There can be issue for data due missing values, lots of outliers, incorrect values, less data, inconsistent data . The ramifications are pretty straight forward, if we us the same data for modelling, our model will take garbage as input and output garbage output with no real meaning.

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

Answer: There are many approaches for categorical data exploration like arbitrary values imputation, frequent category imputation, adding an additional variable to capture Nan.

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

Answer: Missing values or null values can cause complication in data handling and analysis, loss of information and can easily produce biased results from ML model we designed. Some of the simple methods include replacing the missing numerical values like mean, median or by frequently occurring value.

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

Answer: As per my knowledge there are 7 methods of dealing with missing values, whatever their data types are:

1. Deleting rows with missing values but their number should be small.
2. Impute some other values like mean, median or most frequently occurring value.
3. Using algorithms that are not affected by missing values.
4. Using a flag feature to mark the missing value for a particular existing feature in the dataset.

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

Answer: The 4 most important data pre-processing techniques are data cleaning, data integration, data transformation and data reduction.

Sometimes our model performs way better with good quality data and with more columns with same quality. This happens because our model gets confused to find the best fit line/hyperplane for cregression/classification tasks. To make our model better we remove some of our columns since our model was performing better and the process of removing unnecessary columns is dimensionality reduction. Most common dimensionlity reduction techniques are PCA and t-SNE.

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

Answer: IQR is the difference between any feature’s calculated 75th percentile and 25th percentile.

IQR = 75th percentile – 25th percentile.

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: Following are the components of a boxplot:

1. 25th percentile,75th percentile and 50th percentile in the box region
2. IQR = 75th percentile – 25th percentile
3. Lower whisker = 25th percentile value – 1.5\*IQR
4. Upper whisker = 75th percentile +1.5\*IQR
5. Any below above upper whisker and below lower whisker are caller outliers .

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

1. Data collected at regular intervals

2. The gap between the quartiles

Answer The gap between quantiles is known as inter quantile range (IQR) and depicts that most values of the values lie between 25th and 75th percentile value calculated.

3. Use a cross-tab

Answer: Cross tab

1. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

Answer: Histogram is curve which created based on the number of counts for a particular value which is present in the dataset and count of values is done for values lying a certain range decided based on the number of bins.

Box plot is a pictorial representation in more of a statistical manner with outliers, 25th percentiles, median, 75th percentile and IQR.

3. The average and median

Answer: Average is simply the sum of all the value present and is divided by number of values present in your dataset and is affected by outliers and no sorting is done on the values.

For median , the values are first sorted and the elements in the middle are selected for the main median value with no explicit calculation done for the same. Since sorting is done, the main selected value is robust to outliers.