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

Answers:

Key Tasks:

1. Data gathering
2. Data pre-processing
3. Exploratory data analysis (EDA)
4. Feature engineering
5. Training machine learning models of the following kinds:
   1. Regression
   2. Classification
   3. Clustering
6. Multivariate querying
7. Density estimation
8. Dimensionality reduction
9. Model / Algorithm selection
10. Testing and matching
11. Model monitoring
12. Model retraining

Data Pre-processing:

Steps in Data Pre-processing in Machine Learning

1. Acquire the dataset
2. Import all the crucial libraries
3. Import the dataset
4. Identifying and handling the missing values
5. Encoding the categorical data
6. Splitting the dataset
7. Feature scaling

Before starting to train the models, it is of utmost important to prepare data appropriately. As part of data pre-processing, some of the following is done:

* Data cleaning: Data cleaning requires one to identify attributes having not enough data or attributes which are not having variance. These data (rows and columns) need to be removed from training data set.
* Missing data imputation: Handling missing data using data imputation techniques such as replacing missing data with mean, median or mode.

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

Answers: Quantitative data can be counted, measured, and expressed using numbers. Qualitative data is descriptive and conceptual. Qualitative data can be categorized based on traits and characteristics.

Qualitative data is non-statistical and is typically unstructured or semi-structured. This data isn’t necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers. Qualitative data can be used to ask the question “why.” It is investigative and is often open-ended until further research is conducted. Generating this data from qualitative research is used for theorizations, interpretations, developing hypotheses, and initial understandings.

Qualitative data can be generated through:

* Texts and documents
* Audio and video recordings
* Interview transcripts and focus groups
* Observations and notes

Surprisingly enough, identification numbers like a social security number (SSN) or driver’s license are also considered qualitative because they're categorical and unique to one person.

Quantitative data: Contrary to qualitative data, quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended. It can be used to ask the questions “how much” or “how many,” followed by conclusive information

Quantitative data can be generated through:

* Tests
* Experiments
* Surveys
* Market reports
* Metrics

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

Answers: df = (1, 2.0, “Text”)

Integer = 1

Float =2

String = “Text”

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

Answers:

1. Data Collection. Data plays a key role in any use case.
2. Less Amount of Training Data.
3. Non-representative Training Data.
4. Poor Quality of Data.
5. Irrelevant/Unwanted Features.
6. Overfitting the Training Data.
7. Underfitting the Training data.
8. Offline Learning & Deployment of the model.
9. Understanding Which Processes Need Automation.
10. Lack of Quality Data.
11. Inadequate Infrastructure.
12. Implementation.
13. Lack of Skilled Resources.

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

Answers:

Categorical variables take on a fixed and limited number of possible values. For example – grades, gender, blood group type, etc. Also, in the case of categorical variables, the logical order is not the same as categorical data e.g. “one”, “two”, “three”. But the sorting of these variables uses logical order. For example, gender is a categorical variable and has categories – male and female and there is no intrinsic ordering to the categories. A purely categorical variable is one that simply allows you to assign categories, but you cannot clearly order the variables.

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

Answers: Missing values in datasets can cause the complication in data handling and analysis, loss of information and efficiency, and can produce biased results. We can drop the data with missing values or impute them with mean, median, or most frequently occurring values or by other statistical methods.

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

Answers:

Various methods for dealing with missing data values in depth:

1. Deleting Rows: This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.
2. Replacing With Mean/Median/Mode: This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.

1. Assigning An Unique Category: A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values. Here, the features Cabin and Embarked have missing values which can be replaced with a new category, say, U for ‘unknown’. This strategy will add more information into the dataset which will result in the change of variance. Since they are categorical, we need to find one hot encoding to convert it to a numeric form for the algorithm to understand it.
2. Predicting The Missing Values: Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. We will be using linear regression to replace the nulls in the feature ‘age’, using other available features. One can experiment with different algorithms and check which gives the best accuracy instead of sticking to a single algorithm.
3. Using Algorithms Which Support Missing Values: KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values. In this particular dataset, taking into account the person’s age, sex, class etc, we will assume that people having same data for the above mentioned features will have the same kind of fare.

Unfortunately, the SciKit Learn library for the K – Nearest Neighbour algorithm in Python does not support the presence of the missing values.

Another algorithm which can be used here is RandomForest. This model produces a robust result because it works well on non-linear and the categorical data. It adapts to the data structure taking into consideration of the high variance or the bias, producing better results on large datasets.

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

Answers:

Data Pre-processing in Machine Learning:

1. Import libraries

2. Read data

3. Checking for missing values

4. Checking for categorical data

5. Standardize the data

6. PCA transformation

7. Data splitting

Before starting to train the models, it is of utmost important to prepare data appropriately. As part of data pre-processing, some of the following is done:

* Data cleaning: Data cleaning requires one to identify attributes having not enough data or attributes which are not having variance. These data (rows and columns) need to be removed from training data set.
* Missing data imputation: Handling missing data using data imputation techniques such as replacing missing data with mean, median or mode.

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

Feature Selection

* Remove features with missing values
* Remove features with low variance
* Remove highly correlated features
* Univariate feature selection
* Recursive feature elimination
* Feature selection using SelectFromModel

Dimensionality Reduction

* PCA

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?

Answers:

1. The interquartile range shows how the data is spread about the median. The median is the corresponding measure of central tendency. The IQR can be used to identify outliers.
2. A box plot (aka box and whisker plot) uses boxes and lines to depict the distributions of one or more groups of numeric data. Box limits indicate the range of the central 50% of the data, with a central line marking the median value. Lines extend from each box to capture the range of the remaining data, with dots placed past the line edges to indicate outliers.

Construction of a box plot is based around a dataset’s quartiles, or the values that divide the dataset into equal fourths. The first quartile (Q1) is greater than 25% of the data and less than the other 75%. The second quartile (Q2) sits in the middle, dividing the data in half. Q2 is also known as the median. The third quartile (Q3) is larger than 75% of the data, and smaller than the remaining 25%. In a box and whiskers plot, the ends of the box and its center line mark the locations of these three quartiles.

The distance between Q3 and Q1 is known as the interquartile range (IQR) and plays a major part in how long the whiskers extending from the box are. Each whisker extends to the furthest data point in each wing that is within 1.5 times the IQR. Any data point further than that distance is considered an outlier, and is marked with a dot.

When a data distribution is symmetric, you can expect the median to be in the exact center of the box: the distance between Q1 and Q2 should be the same as between Q2 and Q3. Outliers should be evenly present on either side of the box. If a distribution is skewed, then the median will not be in the middle of the box, and instead off to the side. You may also find an imbalance in the whisker lengths, where one side is short with no outliers, and the other has a long tail with many more outliers.

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

Answers:

1. Interval data is a type of data which is measured along a scale, in which each point is placed at an equal distance (interval) from one another. Unlike ordinal data, interval data always take numerical values where the distance between two points on the scale is standardised and equal.

3. Cross tab is used to summarize data related, to determine if there is an association between two variables measured at the nominal or ordinal levels, we use cross-tabulation and a set of supporting statistics. A cross-tabulation (or just crosstab) is a table that looks at the distribution of two variables simultaneously.

11. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median

Answers:

1. Nominal data is a group of non-parametric variables, while Ordinal data is a group of non-parametric ordered variables. Although, they are both non-parametric variables, what differentiates them is the fact that ordinal data is placed into some kind of order by their position

2. Histograms and box plots are graphical representations for the frequency of numeric data values. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets.

3. The average is the arithmetic mean of a set of numbers. The median is a numeric value that separates the higher half of a set from the lower half.