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

* **Data preprocessing is the process of transforming raw data into an understandable format**

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

|  |  |
| --- | --- |
| **Quantitative** | **Qualitative** |
| * **Quantitative data are used when a researcher is trying to quantify a problem, or address the "what" or "how many" aspects of a research question.** * **It is data that can either be counted or compared on a numeric scale.** * **For example, it could be the number of first year students at Macalester, or the ratings on a scale of 1-4 of the quality of food served at Cafe Mac.** * **This data are usually gathered using instruments, such as a questionnaire which includes a ratings scale or a thermometer to collect weather data.** * **Statistical analysis software, such as SPSS, is often used to analyze quantitative data.** | * **Qualitative data describes qualities or characteristics.** * **It is collected using questionnaires, interviews, or observation, and frequently appears in narrative form.** * **For example, it could be notes taken during a focus group on the quality of the food at Cafe Mac, or responses from an open-ended questionnaire. Qualitative data may be difficult to precisely measure and analyze. The data may be in the form of descriptive words that can be examined for patterns or meaning, sometimes through the use of coding.** * **Coding allows the researcher to categorize qualitative data to identify themes that correspond.** |

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

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

* **Inadequate Training Data.**
* **Poor quality of data.**
* **Non-representative training data.**
* **Overfitting and Underfitting**
* **Monitoring and maintenance.**
* **Getting bad recommendations.**
* **Lack of skilled resources.**
* **Customer Segmentation**

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

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

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

* **Replacing With Mean if the data is in numeric form**

**This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.**

* **Replacing With Mode**

**if the data is in categoricle form**

* **Replacing With Median**

**Median is the middlemost value. It’s better to use the median value for imputation in the case of outliers.**

* **If a row has many missing values then you can choose to drop the entire row.**
* **If every row has some (column) value missing then you might end up deleting the whole data.**
* **If you can make an educated guess about the missing value then you can replace it with some arbitrary value using the following code**

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

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?

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