1. What are the key tasks involved in getting ready to work with machine learning modeling?

Ans:

* Collecting Data
* Preparing the Data
* Choosing a Model
* Training the Model
* Evaluating the Model
* Parameter Tuning
* Making Predictions

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

Ans:

**DATA:** It can be any unprocessed fact, value, text, sound, or picture

**DATA:** It can be Qualitative (Nominal,Ordinal) or Quantitative (Discrete, Continuous)

Ex Nominal : Gender ( Women,Men)

Ordinal : First,,Second, Third

Letter grade (A,B,C)

Economic status (Low,High)

Discrete: The number of students in class

The number workers in a company

Continuous : The height of children

The sq footage of a two-bedroom house

**DATA:** it can be Numerical, Categorical, Times series, Text data

3. Distinguish:

**1. Numeric vs. categorical attributes**

Definitions

Categorical data is a type of data that is used to group information with similar characteristics while Numerical data is a type of data that expresses information in the form of numbers. It combines numeric values to depict relevant information while categorical data uses a descriptive approach to express information

Other Names

Categorical data is also called qualitative data while numerical data is also called quantitative data. This is because categorical data is used to qualify information before classifying them according to their similarities.

Numerical data is used to express quantitative values and can also perform arithmetic operations which is a quantitative characteristic.

Types

Categorical data is divided into two types, namely; nominal and ordinal data while numerical data is categorised into discrete and continuous data. Continuous data is now further divided into interval data and ratio data.

Data Characteristics

The characteristics of categorical data include; lack of a standardized order scale, natural language description, takes numeric values with qualitative properties, and visualized using bar chart and pie chart.

Numerical data, on the other hand, has a standardized order scale, numerical description, takes numeric values with numerical properties, and is visualized using bar charts, pie charts, scatter plots, etc.

User-centred Design

Numerical data collection method is more user-centred than categorical data. Most respondents do not want to spend a lot of time filling out forms or surveys which is why questionnaires used to collect numerical data have a lower abandonment rate compared to that of categorical data.

This is because categorical data is mostly collected using open-ended questions.

Data Collection Methods

Categorical data can be collected through different methods, which may differ from categorical data types. For instance, nominal data is mostly collected using open-ended questions while ordinal data is mostly collected using multiple-choice questions.

Numerical data, on the other hand, is mostly collected through multiple-choice questions. We observe that it is mostly collected using open-ended questions whenever there is a need for calculation.

Data Collection Tools

Data collectors and researchers collect numerical data using questionnaires, surveys, interviews, focus groups and observations. Categorical data is collected using questionnaires, surveys, and interviews.

Data collection is usually straightforward with categorical data and hence, does not require technical tools like numerical data. For example, numerical data of a participant’s score in different sections of an IQ test may be required to calculate the participant’s IQ.

When collected using online forms, this may require some technical additions to the form, unlike categorical data which is simple.

Analysis & Interpretation

There are 2 methods of performing numerical data analysis, namely; descriptive and inferential statistics. Some examples of these 2 methods include; measures of central tendency, turf analysis, text analysis, conjoint analysis, trend analysis, etc.

There are also 2 methods of analyzing categorical data, namely; median and mode. In some cases, we see that ordinal data Is analyzed using univariate statistics, bivariate statistics, regression analysis, etc. which is used as an alternative to calculating mean and standard deviation.

Uses

Numerical data is mostly used for calculation problems in statistics due to its ability to perform arithmetic operations. For example, when designing a CGPA calculator, one may need to include commands that allow for the addition, subtraction, division, and multiplication.

Categorical data, on the other hand, is mostly used for performing research that requires the use of respondent’s personal information, opinion, etc. It is commonly used in business research.

Advantage

Numerical data is compatible with most statistical analysis methods and as such makes it the most used among researchers. Categorical data, on the other hand, does not support most statistical analysis methods.

There are alternatives to some of the statistical analysis methods not supported by categorical data. However, they can not give results that are as accurate as the original.

Disadvantage

Numerical data analysis is mostly performed in a standardized or controlled environment, which may hinder a proper investigation. This is because natural factors that may influence the results have been eliminated, causing the results not to be completely accurate.

Numerical data collection is also strictly based on the researcher’s point of view, limiting the respondent’s influence on the result. This is not the case with categorical data.

Nominal data captures human emotions to an extent through open-ended questions. However, the setback with this is that the researcher may sometimes have to deal with irrelevant data.

**2. Feature selection vs. dimensionality reduction**

While both methods are used for reducing the number of features in a dataset, there is an important difference. Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension.

Feature Selection

Feature selection yields a subset of features from the original set of features, which are best representatives of the data. It is an exhaustive search.

-In text data, features might be size of characters or some global features of the text. Feature selection will keep only certain features of those.

-Feature selection is done in the context of an optimization problem.

Dimension Reduction

Dimensionality reduction is generic and only depends on the data and not on what you plan to do with it.

Assuming a classification problem you select the features that will help you classify your data better, while a dimensionality reduction algorithm is unaware of this and just projects the data into a lower dimensionality space. That in turn can work quite well or not for your classification algorithm.

4. Make quick notes on any two of the following:

1. The histogram:

A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

2. Use a scatter plot:

Scatter plots are used in either of the following situations.

* When we have paired numerical data
* When there are multiple values of the dependent variable for a unique value of an independent variable
* In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appear to be related both occur with the exact cause and so on.

3.PCA (Personal Computer Aid)

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

Ans: Data investigation is **a process to understand the nature of data. Investigators can make sense of a real-world issue through data and make evidence-based claims and inferences to propose solutions to a problem.**

Data **allows organizations to more effectively determine the cause of problems**. Data allows organizations to visualize relationships between what is happening in different locations, departments, and systems.Data scientists spend most of their time investigating the data and preparing it for modelling. This helps them to build accurate models.

**The combination of qualitative and quantitative data can also lead to clashes in the philosophical assumptions behind each approach**

Quantitative data is fixed and universal. Qualitative data is subjective and unique. Quantitative research methods are measuring and counting. A qualitative approach is subjective, too. It focuses on describing an action, rather than measuring it.

Quantitative data is expressed in numbers that can be counted or measured, whereas qualitative data is descriptive and conceptual.Qualitative research does not simply help to collect data. It gives a chance to understand the trends and meanings of natural actions. It’s flexible and iterative.

6. What are the various histogram shapes? What exactly are ‘bins'?

And: The different types of a histogram are:

Bell-shaped, Skewed,Random, Uniform histogram, Symmetric histogram, Bimodal histogram.

**Bins**: bins represent ranges of data

histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets".

The bins are usually specified as consecutive, non-overlapping [intervals](https://en.wikipedia.org/wiki/Interval_(mathematics)) of a variable. The bins (intervals) must be adjacent and are often (but not required to be) of equal size.If the bins are of equal size, a bar is drawn over the bin with height proportional to the [frequency](https://en.wikipedia.org/wiki/Frequency_(statistics))—the number of cases in each bin. A histogram may also be [normalized](https://en.wikipedia.org/wiki/Normalization_(statistics)) to display "relative" frequencies showing the proportion of cases that fall into each of several [categories](https://en.wikipedia.org/wiki/Categorization), with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its *area* proportional to the frequency of cases in the bin.The vertical axis is then not the frequency but *frequency density*—the number of cases per unit of the variable on the horizontal axis

7. How do we deal with data outliers?

Ans: ways to deal with outliers in data.

1. Set up a filter in your testing tool.
2. Remove or change outliers during post-test analysis
3. Change the value of outliers.
4. Consider the underlying distribution
5. Consider the value of mild outliers.

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

Ans: Mean, Median and Mode are central inclination measures.

**The Median is the middle score for a set of data that has been arranged in order of magnitude.**

**The Mean equal to sum of all the values in the dataset divided by the number of values in the data.**

Mean is susceptible to influence of outliers. These values are unusual compared to the rest of the data set by being especially small or large in numerical value. The mean is being skewed by two large or small data points values.

Also when our data is skewed, the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from typical value.However, the median best retains this position and is not as strongly influenced by skewed values.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

Ans: A scatter plot is a graph that shows the relationship between data that has two numeric variables. For example, if you collect data on the relationship between the number of hours a student studies and their score on a test, the data has two numeric variables: time studied and test score. You graph scatter plot values on the coordinate plane.

The distance of the points to the line is called "scatter". A large amount of scatter around the line indicates a weak relationship. Little scatter represents a strong relationship. If all points fall directly on a straight line, we have a perfect linear relationship between our two variables.

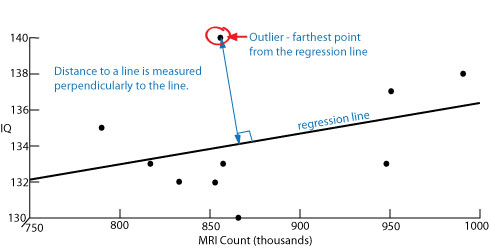
**We often see patterns or relationships in scatterplots.**

When the y variable tends to increase as the x variable increases, we say there is a positive correlation between the variables.

Finding outliers using scatter plot:

If there is a regression line on a scatter plot, you can identify outliers. An outlier for a scatter plot is the point or points that are farthest from the regression line. There is at least one outlier on a scatter plot in most cases, and there is usually only one outlier. Note that outliers for a scatter plot are very different from outliers for a boxplot.

Distance from a point to the regression line is the length of the line segment that is perpendicular to the regression line and extends from the point to the regression line. If one point of a scatter plot is farther from the regression line than some other point, then the scatter plot has at least one outlier. If a number of points are the same farthest distance from the regression line, then all these points are outliers. If all points of the scatter plot are the same distance from the regression line, then there is no outlier.

****

10. Describe how cross-tabs can be used to figure out how two variables are related.

Ans: By showing how correlations change from one group of variables to another, cross tabulation allows for the identification of patterns, trends, and probabilities within data sets. When it comes to analyzing survey response data, cross tabulation reports depict the relationship between two or more survey questions.

A cross-tabulation (or just crosstab) is a table that looks at **the distribution of two variables** simultaneously.

A cross tabulation groups variables to understand the correlation between different variables.

Use crosstab contingency tables with chi-square analysis for categorical data. The test helps you determine if **two discrete variables a**re associated.