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

Machine Learning Steps

* 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.

Data can come in many forms, but machine learning models rely on four primary data types. These include numerical data, categorical data, time series data, and text data.

Training data is the initial dataset used to train machine learning algorithms. Models create and refine their rules using this data. It's a set of data samples used to fit the parameters of a machine learning model to training it by example

Data can be defined by a set of variables with qualitative or quantitative nature.

three shapes of data: structured data, unstructured data, and semi-structured data.

STRUCTURED DATA

Structured data is tabular data, containing columns and rows which are very well defined. The main advantage of this type of data is being easily stored, entered, queried, modified, and analyzed. Structured data is often managed by Structured Query Language, or SQL - a programming language created for managing and querying data in relational management systems.

UNSTRUCTURED DATA

Unstructured data is the rawest form of any data, and it can be in any type or file: pictures and graphic images, webpages, PDF files, videos, emails, word processing documents, etc. This data is often stored in repositories of files. Extracting valuable information out of this type of data can be somewhat challenging. For example, a text can be analyzed by extracting the topics it covers and whether the text is positive or negative about them.

SEMI-STRUCTURED DATA

As the name implies, semi-structured data is a cross between structured and unstructured data. A semi-structured data may have a consistent defined format, however, the structure may not be very strict. The structure may not be necessarily tabular and parts of the data may be incomplete, or contain differing types. An example can be photos of other graphics tagged with keywords, making it easy to organize and locate graphics.

HISTORICAL AND REAL-TIME DATA

Historical datasets can help in answering exactly the types of questions that decision-makers would like to benchmark against real-time data. Historical data sources can be best suited for building or modifying predictive or prescriptive models, and offering insights that can improve long-term and strategic decision making. The basic definition or real-time data explains it as a data that is passed along the end-user as quickly as it is gathered. Real-time data can be enormously valuable in things like traffic GPS systems, in bench-marking different kinds of analytics projects and for keeping people informed through instant data delivery.

In predictive analytics, both types of data sources should be given equal consideration, as both can help in predicting and identifying future trends.

INTERNAL AND EXTERNAL DATA

INTERNAL DATA

Internal data is information gathered within an organization and can cover areas such as personnel, operations, finance, maintenance, procurement, and many more. Internal data can provide information on employee turnover, sales success, profit margins, structure and dynamics of an organization, etc.

EXTERNAL DATA

External data is the information gathered from outside, including customers, staging websites, agencies, and more. For example, external data gathered from social media can provide insights about the behavior, preferences, and motivations of customers. At this stage, you may wonder if internal data is the same as primary data, and external data the same as secondary data. This is close but slightly different. The categorization of internal and external data sources is mostly in terms of where the data comes from - whether it was collected from your organization or from a source outside your organization. The notion of primary/secondary data rather refers to the purpose and time-frame for which the data was collected - whether it was collected by the researcher for a precise project, or in the form of another source, even within the same organization.

These different types of data sets can be found within an organization, but they can also be found in external data sources like the internet. We help companies make smarter and data-driven decisions with the help of Artificial Intelligence and Machine Learning. If you are curious to find out how your organization can leverage data to boost growth, don't hesitate to contacts us, and one of our team members will be back at you shortly with more insights!

3. Distinguish:

1. Numeric vs. categorical attributes

In the machine learning world, data is nearly always split into two groups: numerical and categorical. Numerical data is used to mean anything represented by numbers (floating point or integer). Categorical data generally means everything else and in particular discrete labeled groups are often called out.

Categorical data refers to a data type that can be stored and identified based on the names or labels given to them. Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form. Also known as qualitative data as it qualifies data before classifying it.

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.

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

1. The histogram

A histogram is a graph used to represent the frequency distribution of a few data points of one variable. Histograms often classify data into various “bins” or “range groups” and count how many data points belong to each of those bins.

2. Use a scatter plot

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

3.PCA (Personal Computer Aid)

Computer-aided or computer-assisted is an adjectival phrase that hints of the use of a computer as an indispensable tool in a certain field, usually derived from more traditional fields of science and engineering.

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

Good data provides indisputable evidence, while anecdotal evidence, assumptions, or abstract observation might lead to wasted resources due to taking action based on an incorrect conclusion.

It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

The combination of qualitative and quantitative data can also lead to clashes in the philosophical assumptions behind each approach and therefore recommendations have been made for triangulation to be carried out from a pragmatic, or subtle realist, approach

Simply put, quantitative data gets you the numbers to prove the broad general points of your research. Qualitative data brings you the details and the depth to understand their full implications. To get the best results from these methods in your surveys, it's important that you understand the differences between them.

Quantitative research methods are measuring and counting. Qualitative research methods are interviewing and observing. Quantitative data is analyzed using statistical analysis. Qualitative data is analyzed by grouping the data into categories and themes.

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

* Skewed Distribution.
* Double-Peaked or Bimodal
* Plateau or Multimodal Distribution
* Edge Peak Distribution
* Comb Distribution
* Truncated or Heart-Cut Distribution
* Dog Food Distribution.

A 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".

7. How do we deal with data outliers?

1. Set up a filter in your testing tool

Even though this has a little cost, filtering out outliers is worth it. You often discover significant effects that are simply “hidden” by outliers.

2. Remove or change outliers during post-test analysis

One way to account for this is simply to remove outliers, or trim your data set to exclude as many as you’d like.

3. Change the value of outliers

Essentially, instead of removing outliers from the data, you change their values to something more representative of your data set. It’s a small but important distinction: When you trim data, the extreme values are discarded.

4. Consider the underlying distribution

Traditional methods to calculate confidence intervals assume that the data follows a normal distribution, but as with certain metrics like average revenue per visitor, that usually isn’t the way reality works.

5. Consider the value of mild outliers

There’s a chance that, in your data analysis, you shouldn’t throw away outliers. Rather, you should segment them and analyze them more deeply

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

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency:

* mode
* median
* mean

Each of these measures describes a different indication of the typical or central value in the distribution.

Mode

The mode is the most commonly occurring value in a distribution.

Median

The median is the middle value in distribution when the values are arranged in ascending or descending order.

Mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

Mean may vary too much from median in certain data sets when the data set has a lot of values in one particular side of the distribution. If mean and median are considerably different, this indicates that the data are skewed (i.e. they are far from being normally distributed) and the median generally gives a more appropriate idea of the data distribution.

The mean reflects the skewing the most. The mean is affected by outliers that do not influence the mean. Therefore, when the distribution of data is skewed to the left, the mean is often less than the median. When the distribution is skewed to the right, the mean is often greater than the median

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

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.

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.

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

Cross tabulation is used to quantitatively analyze the relationship between multiple variables. Cross tabulations — also referred to as contingency tables or crosstabs — group variables together and enable researchers to understand the correlation between the different variables.

It identifies patterns, trends, and association between variables which could otherwise be difficult to understand and interpret.

A cross tabulation (or crosstab) report is used to analyze the relationship between two or more variables. The report has the x-axis as one variable (or question) and the y-axis as another variable. This type of analysis is crucial in finding underlying relationships within your survey results. (or any type of data!)

Cross tabulations — also referred to as contingency tables or crosstabs — group variables together and enable researchers to understand the correlation between the different variables.

#1. Cross tabulation helps to reduce confusion while analyzing data.

Large data sets can be overwhelming and confusing. Pulling insights from them to inform business decisions can often be a daunting task.

By creating crosstabs, data sets are simplified by dividing the total set into representative subgroups, which can then be interpreted at a smaller, more manageable scale.

This reduces the potential for making errors while analyzing the data, which means that time is spent efficiently.

#2. Cross tabulation allows for profound data insights.

By reducing total data sets into more manageable subgroups, cross tabulation allows researchers to yield more granular, profound insights.

The insights into the relationships between categorical variables resulting from cross tabulation would be impossible to obtain by digging into the set as a whole. This means that if crosstabs were not created, these insights would go unnoticed, or at the very least they’d require much more legwork to reveal.

#3. Insights derived from cross tabulation are actionable.

The entire purpose of performing statistical analysis on a data set is to uncover actionable insights that will then impact your end goal. These insights are able to impact business by backing up thought processes and decision-making with hard data.

Because cross tabulation simplifies complex data sets, these impactful insights are much easier to expose, record, and consider while developing overarching strategies.