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

Step 1: Collect Data. ...

Step 2: Prepare the data. ...

Step 3: Choose the model. ...

Step 4 Train your machine model. ...

Step 5: Evaluation. ...

Step 6: Parameter Tuning. ...

Step 7: Prediction or Inference

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

Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text

3. Distinguish:

1. Numeric vs. categorical attributes

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.

1. 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 the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar.

* The horizontal axis displays the number range.
* The vertical axis (frequency) represents the amount of data that is present in each range.

The number ranges depend upon the data that is being used.

1. Use a scatter plot

scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.

3.PCA (Personal Computer Aid)

The principal application of PCA is dimension reduction. If you have high dimensional data, PCA allows you to reduce the dimensionality of your data so the bulk of the variation that exists in your data across many high dimensions is captured in fewer dimensions

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

Collecting data allows you to store and analyze important information about your existing and potential customers. Collecting this information can also save your company money by building a database of customers for future marketing and retargeting efforts.

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.

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

Different shape of Histogram as under:

Frequency Distributions

Probability Distributions

The most common distribution shapes are

Symmetric

Bell-shaped

Skewed to the left

Skewed to the right

Uniform

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?

5 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?

Mean, Median & Mode

Due to Different types of Data Distribution

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

Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation . ... The closer the data points come when plotted to making a straight line, the higher the correlation between the two variables, or the stronger the relationship.

Yes, we can find out using scatter plot

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

Cross tabulation is a method to quantitatively analyze the relationship between multiple variables. Also known as contingency tables or cross tabs, cross tabulation groups variables to understand the correlation between different variables. It also shows how correlations change from one variable grouping to another.