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

Answer: The 7 Key Steps to Build Your Machine Learning Model

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.

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

1.) Numerical data:

* Continuous data: Height, Weight, salary, temperatures, interest rates.
* Discrete data: units sold, number of languages spoken, number of students.

2.) Categorical data: Categorical data represents characteristics, such as a hockey player’s position, team, hometown. Categorical data can take numerical values. For example, maybe we would use 1 for the colour red and 2 for blue. But these numbers don’t have a mathematical meaning. That is, we can’t add them together or take the average.

3.) Time-series data: Time series data is a sequence of numbers collected at regular intervals over some period of time. For example, we might measure the average number of home sales for many years.

4.) Text: Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction

Answer:

1. 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.
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

* 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

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

1. The histogram

2. Use a scatter plot

3.PCA (Personal Computer Aid)

Answer:

1. Histogram: A histogram is a graphical representation that organizes a group of data points 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. Scatter Plot: A scatter plot 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.

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

Answer: If the data set is messy, building models will not help to solve the problem. What will happen is “garbage in, garbage out.” In order to build a powerful machine learning algorithm. We need to explore and understand our data set before we define a predictive task and solve it. Before going further, Data scientists spend most of their time exploring, cleaning, and preparing their data for modeling. This helps them to build accurate models and check assumptions required for fitting models. Create meaningful data visualizations, predict future trends from the data. If we are good at understanding data preparation almost 80% of the work is completed.

Quantitative and qualitative research both have their place in market research and a mix of both should be carried out whenever you’re extending product lines or launching something new. Both methods can work hand-in-hand; brands can use qualitative research for developing concepts and theories, and quantitative for testing pre-existing ones. You can also use free-form qualitative research to guide the creation of more structured qualitative surveys. And following quantitative surveys, turn to qualitative to better understand the context of the responses!

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

Answer: A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars. Each bar typically covers a range of numeric values called a bin or class; a bar's height indicates the frequency of data points with a value within the corresponding bin.

7. How do we deal with data outliers?

Answer:

5 ways to deal with outliers in data:

1. Set up a filter in your testing tool. Even though this has a little cost, filtering out outliers is worth it.
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?

Answer: There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.

The outlier does not affect the median. The outlier decreases the mean so that the mean is a bit too low to be a representative measure of this student's typical performance. This makes sense because when we calculate the mean, we first add the scores together, then divide by the number of scores.

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

Answer: The scatter plot is a fundamental tool for looking at bivariate data. It shows the important characteristics of the data and can be used to decide what model may describe the relationship between the variables. Scatter plots and box plots are the most preferred visualization tools to detect outliers. Scatter plots — Scatter plots can be used to explicitly detect when a dataset or particular feature contains outliers.

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

Answer: 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.