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.

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

There are perhaps 14 types of learning that you must be familiar with as a machine learning practitioner; they are:

**Learning Problems**

* 1. Supervised Learning
* 2. Unsupervised Learning
* 3. Reinforcement Learning

**Hybrid Learning Problems**

* 4. Semi-Supervised Learning
* 5. Self-Supervised Learning
* 6. Multi-Instance Learning

**Statistical Inference**

* 7. Inductive Learning
* 8. Deductive Inference
* 9. Transductive Learning

**Learning Techniques**

* 10. Multi-Task Learning
* 11. Active Learning
* 12. Online Learning
* 13. Transfer Learning
* 14. Ensemble Learning

3. Distinguish:

* 1. Numeric vs. categorical attributes

Both numerical and categorical data can take numerical values. Categorical data can take values like identification number, postal code, phone number, etc. The only difference is that **arithmetic operations cannot be performed** on the values taken by categorical data.

* 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 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. Use a scatter plot: Use a scatter plot **to determine whether or not two variables have a relationship or correlation**. Are you trying to see if your two variables might mean something when put together? Plotting a scattergram with your data points can help you to determine whether there's a potential relationship between them.

3.PCA (Personal Computer Aid): A **personal computer** (**PC**) is a multi-purpose **computer** whose size, capabilities, and price make it feasible for **individual** use.

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

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.

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

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?

1. Ans: 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?

Measures of central tendency help you find the middle, or the average, of a data set. ... The median is the middle number in an ordered data set. **The mean is the sum of all values divided by the total number of values**.

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

Each point on the plot shows the X and Y scores for a single subject. This is what we mean by "bivariate" plot -- each point represents two variables. ... **The distance of the points to the line is called** "scatter". A large amount of scatter around the line indicates a weak relationship

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**. ... It also shows how correlations change from one variable grouping to another. It is usually used in statistical analysis to find patterns, trends, and probabilities within raw data.