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

Answer :

below are key tasks :

1. Data collection  
    Defining the problem and assembling a dataset
2. Data preparation  
    Preparing your data
3. Choose model
4. Train model  
    Developing a model that does better than a baseline
5. Evaluate model  
    Choosing a measure of success   
    Deciding on an evaluation protocol
6. Parameter tuning Regularizing your model and tuning your parameters
7. Predict

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

Answer : Data can be categorized into 4 basic types from Machine Learning perspectives :

1. Numerical
2. Categorical
3. TimeSeries
4. Text

1. Numerical : Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data. This data has meaning as a **measurement** such as house prices.

Numerical data can be characterized by continuous or discrete data.

Continuous data can assume any value within a range whereas discrete data has distinct values.

For example, the number of students taking Python class would be a discrete data set. You can only have discrete whole number values like 10, 25, or 33. A class cannot have 12.75 students enrolled. A student either join a class or he doesn’t. On the other hand, continuous data are numbers that can fall anywhere within a range. Like a student could have an average score of 88.25 which falls between 0 and 100.

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.

In the context of super classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.

There is also something called ordinal data, which in some sense is a mix of numerical and categorical data. In ordinal data, the data still falls into categories, but those categories are ordered or ranked in some particular way. An example would be class difficulty, such as beginner, intermediate, and advanced. Those three types of classes would be a way that we could label the classes, and they have a natural order in increasing difficulty.

3. Time Series Data

Time series data is a sequence of numbers collected at regular intervals over some period of time. It is very important, especially in particular fields like finance. Time series data has a temporal value attached to it, so this would be something like a date or a timestamp that you can look for trends in time.

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 data

Answer

Numeric data

1. Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form.

2. Also known as quantitative data as it represents quantitative values to perform arithmetic operations on them.

3. Numeric data have two types : Discrete data and Continuous data.

4. Can be visualized using bar graphs, pie charts as well as scatter plots.

1. Example . score : 80, 90 , height : 5.10

Categorical data :

1. Categorical data refers to a data type that can be stored and identified based on the names or labels given to them.
2. Also known as qualitative data as it qualifies data before classifying it.
3. Categorical data have two types : Nominal data and Ordinal data.
4. Can be visualized using only bar graphs and pie charts.
5. Example : Gender : male/female

2. Feature selection vs. dimensionality reduction

Often, feature selection and dimensionality reduction are grouped together (like here in this article). 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 :

A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to show frequency distributions.

Use a histogram when:

* The data are numerical
* You want to see the shape of the data’s distribution, especially when determining whether the output of a process is distributed approximately normally
* Analyzing whether a process can meet the customer’s requirements
* Analyzing what the output from a supplier’s process looks like
* Seeing whether a process change has occurred from one time period to another
* Determining whether the outputs of two or more processes are different
* You wish to communicate the distribution of data quickly and easily to others

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.

Scatter plots’ primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Identification of correlational relationships are common with scatter plots. In these cases, we want to know, if we were given a particular horizontal value, what a good prediction would be for the vertical value. You will often see the variable on the horizontal axis denoted an independent variable, and the variable on the vertical axis the dependent variable. Relationships between variables can be described in many ways: positive or negative, strong or weak, linear or nonlinear.

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?

Answer : If your data set is messy, building models will not help you to solve your 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.

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

Answer :

Types of histogram :

1. Bell Shape
2. Uniform
3. Bimodel
4. Multimodel
5. Left Skewed
6. Right skewed
7. Random

Bins :

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?

Answer :

**There are some techniques used to deal with outliers.**

1. Deleting observations.
2. Transforming values.
3. Imputation.
4. Separately treating.
5. Deleting observations. Sometimes it's best to completely remove those records from your dataset to stop them from skewing your analysis.

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

Answer :

Three of the many ways to measure central tendency are **the mean, median and mode.**

**Vary too much from median means data has outlier.**

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

Answer : A scatterplot is a type of data display that shows the relationship between two numerical variables. Each member of the dataset gets plotted as a point whose

(x, y)

(*x*,*y*)

left parenthesis, x, comma, y, right parenthesis

coordinates relates to its values for the two variables.

Yes, we can able to see outlier with the help of scatter plot.

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

Answer :

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.