Assignment 04 Solutions ¶

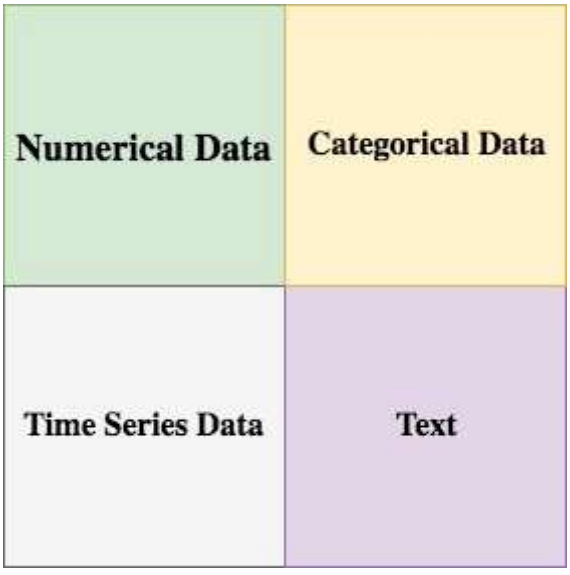
1. What are the key tasks involved in getting ready to work with machine learning modeling ?

Ans: The Key Tasks involved in getting ready to work with Machine learning Modelling are:

1. Data collection: Defining the problem and assembling a dataset.
2. Data preparation: Preparing your data.
3. Choosing a Model
4. Training the Model: Developing a model that does better than a baseline.
5. Evaluating the Model: Choosing a measure of success. Deciding on an evaluation protocol.
6. Parameter tuning: Scaling up: developing a model that overfits.Regularizing your model and tuning your parameters.
7. Prediction or Inference.

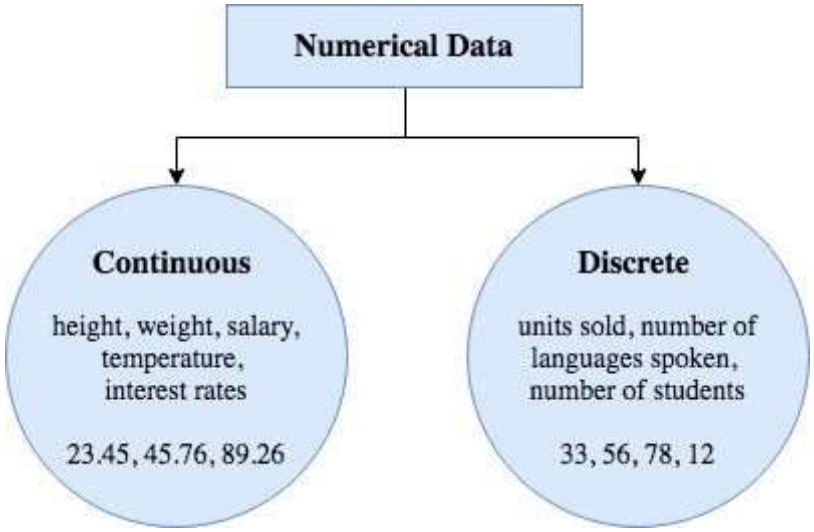
2. What are the different forms of data used in machine learning ? Give a specific example for each of them ?

Ans: Most data can be categorized into 4 basic types from a Machine Learning perspective: `Numerical Data` , `Categorical Data` , `Time-Series Data` , and `Text data` .



Numerical Data:
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 or as a count, such as a number of residential properties in Los Angeles or how many houses sold in the past year.

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.



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.

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.

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 between :

1. Numeric vs. categorical attributes
2. Feature selection vs. dimensionality reduction

Ans: The following are the differences between:

1. **Numeric vs. categorical attributes:**

- Numerical data are values obtained for quantitative variable, and carries a sense of magnitude related to the context of the variable (hence, they are always numbers or symbols carrying a numerical value).
- Categorical data are values obtained for a qualitative variable. categorical data numbers do not carry a sense of magnitude.
- Numerical data always belong to either ordinal, ratio, or interval type, whereas categorical data belong to nominal type. - - Methods used to analyse quantitative data are different from the methods used for categorical data, even if the principles are the same at least the application has significant differences.
- Numerical data are analysed using statistical methods in descriptive statistics, regression, time series and many more.For categorical data usually descriptive methods and graphical methods are employed. Some non-parametric tests are also used.

2. **Feature selection vs. dimensionality reduction**

- Feature selection you just select a subset of the original feature set, without any manipulation of the data on the other hand.
- Dimensionality reduction is typically choosing a new representation within which you can describe most but not all of the variance within your data, thereby retaining the relevant information, while reducing theamount of information necessary to represent it.

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

1. The histogram
2. Use a scatter plot
3. PCA (Personal Computer Aid)

Ans: The Quick notes on the following three topics is:

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.

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.

PCA (Personal Computer Aid): Principal Component Analysis or PCA is a widely used technique for dimensionality reduction of the large data set. Reducing the number of components or features costs some accuracy and on the other hand, it makes the large data set simpler, easy to explore and visualize.

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

Ans: 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' ?

Ans: The different types of a Histogram are:

1. Uniform Histogram
2. Symmetric Histogram
3. Bimodal Histogram
4. Probability Histogram.

The `bin` in a histogram is the choice of unit and spacing on the X-axis. All the data in a probability distribution represented visually by a histogram is filled into the corresponding bins. The height of each bin is a measurement of the frequency with which data appears inside the range of that bin in the distribution.

7. How do we deal with data outliers ?

Ans: We can use `Z-Score` or any of below methods to deal with data outliers:

Univariate Method: This method looks for data points with extreme values on one variable.

Multivariate Method: Here, we look for unusual combinations of all the variables.

Minkowski Error: This method reduces the contribution of potential outliers in the training process.

Z-Score: This can be done with just one line code as we have already calculated the Z-score.

```
boston_df_o = boston_df_o[(z < 3).all(axis=1)]
```

IQR Score: Calculate IQR score to filter out the outliers by keeping only valid values.

```
boston_df_out = boston_df_o1[~((boston_df_o1 < (Q1 - 1.5 * IQR)) |
(boston_df_o1 > (Q3 + 1.5 *IQR))).any(axis=1)]
boston_df_out.shape
```

Quantile function: Use `quantile()` to remove amount of data.

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

Ans: `Mean` , `Median` and `Mode` are Central Inclination Measures. Mean varies more than Median due to presence of outliers, as mean is averaging all points while median in like finding a middle number.

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

Ans: 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. So this visualization gives us the idea of bivariate relationship.

Scatter plot can also help finding outliers as outliers can be visualized at farther distance than regular data.

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

Ans: `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.