2nd may 2022

1

classmate Date Page

Data science chapter 1

- It is a combination of mathematics and statistics, domain knowledge and computer science

- Data: Data are the values of qualitative or quantitative variables belonging to set of item.

set of items may be population we are interested in.
variables are the measurement or characteristics of the item. They might be measured in qualitative terms and quantitative terms.

- Qualitative terms: could be such as country of origin, gender, medical treatment, etc

- Quantitative terms: could be height, weight, blood pressure, etc.

* Data science is applicable of computational & statistical techniques to gain insight into real world problem expressed using data.

Computational because data science typically involves some sort of algorithm, methods which are written inside code.

statistical because inferences based on statistics helps us to build predictions that we make.

- Raw data

Also caued automic data, source data, unprocessed data.

This information may be stored in the file and it can contains collection of numbers, characters, etc.

This data can be entered in computer or it can be generated by computer.

It is hard to parse or analyze.

- Data analysis includes processing or cleaning of data.

Data scientists job is to do processing of such data.

Date
Page

- Processed Data

some kind of cleaning, transformation is perform on raw data to get processed data, which can be analyzed and visualized.

Example: Voltage signal of microphone is raw data and when voltage is filtered and noise is removed, it is processed data.

Processing can include merging, subsetting, transforming, etc.

- Sources of raw data:
  - Binary file generated by measurement machine
  - Unformatted excel file
  - JSON from twitter API
  - Hand entered numbers (readings) you collected.

June 2022

4

Date
Page

- Introduction to univariate, biavariate and multivariate system:

1. Univariate Data

- This type of data consists of only one variable and doesn't deal with a causes or effect relationships then a univariate analysis is used.
- The analysis of univariate data is thus the simplest form of analysis.
- The main purpose of analysis is to describe the data and find patterns that exist within it.

Example: Height

2. Bivariate Data

- Bivariate data slightly more analytical than univariate.
- These type of data involves two different variables and one is dependent on other

Example: Ice-cream sales based on the weather temperature

classmate
Date
Page

3. multivariate Data

- multivariate is more complex form of statistical analysis technique
- Used when there are more than two variable in the data set.

Example: An advertiser want to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationship between these variables can be examined.

- Data Processing:

It occurs when data is collected and translated into useful information.

At initial, the data is unprocessed or raw data. So to use or make that data clean we perform data processing on it. to get processed data.

- six stages of data processing:

Data collection, Data preparation, Data input, Processing, Data Output, Data storage

6

- Data collection

Is the first step of data processing. Is the process of measuring, collecting and analyzing different type of data. using set standard validation technique.

- Data Preparation

Once data is collected, it then enters the data preparation stage.

Data preparation, often referred as "Pre-processing" is the stage at which raw data is cleaned.

Purpose of this step is to eliminate raw data & begin to create high quality data.

- Data visualization and its techniques

Data visualization is graphical representation of information and data. by using visual elements like graphs, charts and maps.

Data visulization convert large and small data sets into visuals, which is easy to understand and process for humans.

Tools & techniques are required to analyze vast amount of information.

7

Data visualization are common in every day life, but they always appear in the form of graphs & charts.

* The term inference and prediction both discribes tasks where we learn from data in a supervised manner in order to find a model that describes the relationship between independent variables and the outcom

- Scatter Plot

Scatter plot uses dots to represent values for two different numeric variables. The position of each dot on horizontal & vertical axis indicates values for an individual data point.

Used to observe relationships between variables.

- Scatter matrix

It is matrix of scatter plots where each scatter plot in grid is created between different combinations of variables.

matrix represents bi-variate or pairwise relationship between different combination of variables while lying them in grid form.

- Histogram

Histogram provides visual representation of the distribution of data set: location, spread and skewness of the data.
- It also helps to visualize whether the distribution is symmetric or skewed left or right.

- Box plot

Also known as five Number summary,
minimum
upper ~~lower~~ quartile Q1
lower quartile Q3
median
maximum

Date
Page

It allows you to see important characteristics of the data visually.

Box plot is a graph that gives you a good indication of how the values in the data spread out.