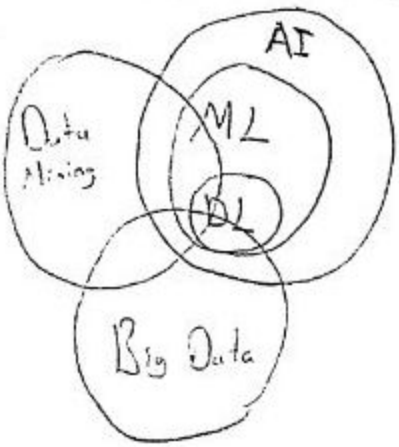
Introduction to Machine Learning - Berk's Part

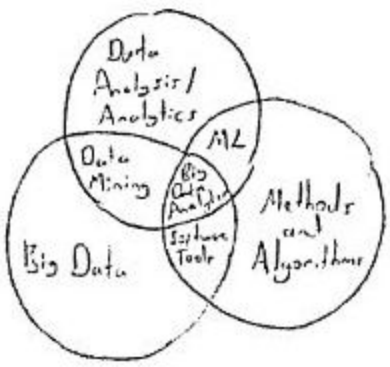
Lecture (1)

(1.1) The Place of ML and AI & DS

Different Parts in Data Science:



The Fields of Data Science:



Big Data: Collecting and processing any data which is huge in volume, arrival/processing rate or invariant in structure

Data Mining: Process of finding out hidden patterns in the structured data and find hidden information in the data.

Data Analytics: It is a process which is one step above data mining. Data analytics identifies the type of the analysis to be performed within which data mining techniques will be performed.

Data Analysis: It is a more general approach of finding insights out of the raw data by forming a hypothesis and proving them using statistical tests.

Data Science: It defines the process of understanding the business problem to deliver the solution.

Machine Learning: It is a tool used in data analytics to predict/find out a hidden layer of information in data.

(1.2) Data Science Steps

Step 1: Collecting Data
- Generally, data is not ready to consume!
- Data comes from many sources such as client database, web logs, etc.
- 3 V Problem (mostly in Big Data)
  - Volume: Data in TB.
  - Velocity: Streaming data with high throughput
  - Veracity: Data with varying structure.

Step 2: Cleaning Data
- Removing discrepancies from data
- Outlier Analysis, Missing Value Reduction, transforming, modelling, visualization
- Data Analysis fits here.

Step 3: Analysing Data
- Create a plan to do analytics on data.
- May include Descriptive Analytics, Predictive Analytics, Prescriptive Analytics
- Data Mining & Data Analytics fit here

Step 4: Drawing Insights & Business Intelligence Reports
- We make future predictions and validate our previously defined hypothesis
- Machine Learning fits here

(1) **Step 5:** Taking Actions!

1.3 The Terms: "Learning" and "Machine Learning"

(Machine) Learning: is programming computers to optimize a performance criterion using example data or past experience.

- Learning general models from a data of particular examples.
- Data is cheap & abundant; knowledge is expensive & scarce

Aim: Build a model that is a good and useful approximation to the data

Machine Learning's sub-roles:

Role of Statistics	Role of Computer Science
- Inference from a sample.	- Solve the optimization problem - Representing and Evaluating the model for inference.

1.4 Data Types

Data: Collection of factual information based on numbers, words, observations, measurements which can be utilized for calculation, discussion and reasoning.

- Data is plural for Datum in Latin Language.
- Data is kind of postulate/axiom in math, they are accepted as base truth, where we don't care some or other.

Categorical/Qualitative Data: Based on descriptive information (e.g, "He is a clever boy")

- **Binomial/Boolean**: Variable data with only 2 options (e.g, good-bad, true-false)
- **Nominal/Unordered**: Variable data which is in unordered form (e.g, red-green-blue, single-married-widow)
- **Ordinal**: Variable data with proper order (e.g, short-med-tall, bad-good-v.good)

Numerical/Quantitative Data: Based on numerical information (e.g, we have 4 lectures)

- **Discrete**: Countable data. (e.g, ID data, # of people)
- **Continuous**: Measureable data. (e.g, height, width, length)
  - **Interval**: No true zero (e.g, absence of temperature, decibel of sounds)
  - **Ratio**: Absolute zero. (e.g, height, duration)

Special Data: Data which is not classifiable like numerical or categorical (Eg, datetime, web url, sound, image)

1.5 Data Representation in Computer world

Main Data Formats: Frequently-used data formats in file-system.

- **csv**: A text format. Comma-seperated-values, column seperator can vary, e.g, ',', '|', '_' characters can be used as delimiter.
- **arff**: A text format. Mainly used for WEKA. Expanded form is Attribute-Relation File Format. Data + Meta-data.
- **office formats**: odt, xlsx. Not in text format, not easy to directly use.

Other Data Formats: Image, sound, video, a database table

JSON, XML

Structured vs. Unstructured Data:

Structured Data: Data where elements are addressable for effective analysis. Simplest way to manage information. (e.g, a relational database (SQL))

Semi-Structured Data: Information that does not reside in a relational database but that have some organizational properties that make it easier to analyze. (e.g, XML data, JSON)

Unstructured Data: A data which is not organized in a pre-defined manner or does not have a pre-defined data model. (e.g, pdf, media logs, office formats)

Meta-data: Data about data. It's compressed

What is 101001?: "101001" or $(101001)_{10}$ or $(101001)_2$

(1.6) Data Table Example & Feature Terms

Obere Dataset:

ID	Height (in cm)	Weight (in lb)	Body Fat Percentage	Age	Gender (F-1, M-0)	Parent's Obesity	isObese
1	165	97	0.26	19	Ø	NO-O	Yes
2	182	126	0.35	41	1	O-O	Yes
3	178	78	0.29	18	1	NO-NO	No
4	182	110	0.30	25	Ø	NO-NO	Yes
5	176	65	0.21	32	2	NO-O	No

(NOTE: INPUT VARIABLE; OUTPUT VARIABLE)

Feature Engineering/Inspection:

ID: Index/Primary-key/Meta-data

Height: Input Variable, Numerical & Discrete

Weight: " , " & "

BFP: Input Variable, Numerical & Continuous & Ratio

Age: Input Variable, Numerical & Discrete

Parent's obesity: Input Variable, Categorical & Ordinal

Gender: Input Variable, Categorical & Binomial (or Nominal if there is other option)

isObese: Output Variable, Categorical & Binomial (also class)

Terms:

- Features = Attributes = Variables = Column
- Input Variable = Independent Variable (if variables are independent from each other, dependency among input variables is discouraged)
- Instance = Observation = Row
- Output Variable = Dependent Variable = Class

Postulate/Axiom of the data:

- Postulates (like in math) are present in data science

Our Main Postulate: IV: Input Variable, OV: Output Variable

There is a function W for every variables, such that

$W(IV_2, IV_1, \ldots, IV_N) = OV$, for every instance

- Our main purpose is to find best W, i.e. the system gives the best output value with the given input values

Without Class: (i.e. Unsupervised Learning)

- Without class, we have only features and can only compare with each other;
- Generally, much more ambiguous and needs more data.

Note That: If no repetition and no ordinality in feature, that feature is useless!

Brainstorming: Why do we use (machine) learning in order to calculate instead of finding a formula like Body-Mass Index?

- Data can change.
- Norm for class (e.g. obesity) can change.
- Data teaches us what the truth is.

1. Extra

Where to find datasets?:

- Kaggle
- GitHub
- Google Datasets
- UCI Machine Learning Repository

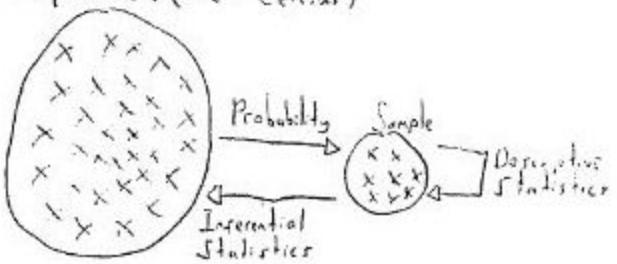
Data Science Tools:

- WEKA
- Scikit Learn
- TensorFlow
- PyTorch

(1.7) Descriptive Statistics

Statistics in a nutshell:

Population (a.k.a Census)



Population Data: Collection of all items of interest

- Denoted by "N".
- The numbers we obtain = PARAMETERS

Subset Data: Subset of population

- Denoted by "n"
- The numbers we obtain = STATISTICS

Frequency Distribution Table: A graphical representation of variables.

Grade	Frequency (Abs)	Frequency (Relative)
Prep School	14	14/94
1	36	36/94
2	32	32/94
3	5	5/94
4	5	5/94
4+	1	1/94
Master/PhD	1	1/94
$\Sigma = 94$		$\Sigma = 94/94 = 1$

Measures of Central Tendency:

- A single value that explains a set of data by identifying the central position within that set of data.
- Also called "Measures of Central Location".
- 3 Measures: Mean, mode, median

Mean: Most popular, used with both discrete & continuous data.

- Sample Mean: $\bar{x} = \frac{\Sigma x}{n}$
- Population Mean: $\mu = \frac{\Sigma x}{N}$

Median:

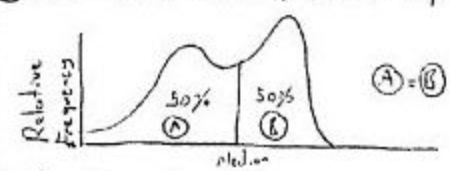
Why we use Median?: Suppose we have a dataset (Yearly Income of Employee at a Company)

18.5 (22.8) (23.7) (24.6) (26.8) (25.2) (122.5) (K$)

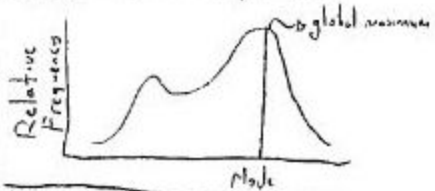
→ $\bar{x} = 47.4$, which is misleading!

→ $\tilde{x} = 26.6$, better "center".

- Resilient to outliers!
- Sample Median: $\tilde{x}$ Population Median: Md
- Median divides area of Relative Frequency into 2.

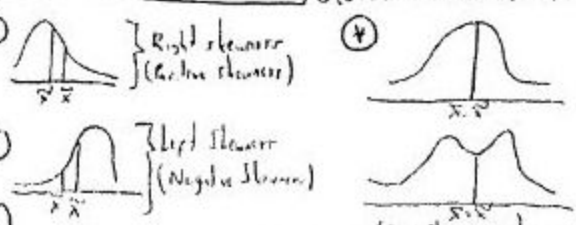


Mode: Mo, most frequent item



Measures of Skewness (a.k.a Measures of Asymmetry)

- Right skewness (Positive skewness)
- Left skewness (Negative skewness)
- (No skewness)



Measures of Variability : The range, the variance, the standard deviation

Range : The simplest method

$$R = x_{max} - x_{min}$$

Variance :

For sample : $s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$

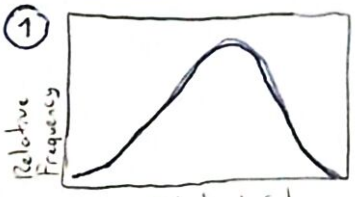
For population : $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$

Standard Deviation :

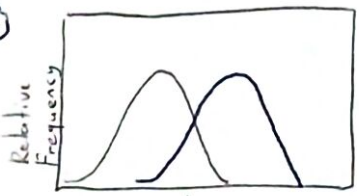
For sample : $s = \sqrt{\frac{\sum (x - \bar{x})}{n-1}}$

For population : $\sigma = \sqrt{\frac{\sum (x - \mu)}{N}}$

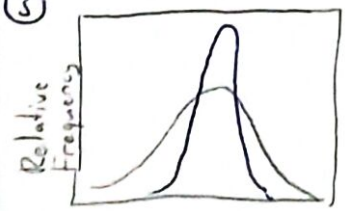
Difference Between Two Datasets :

① 

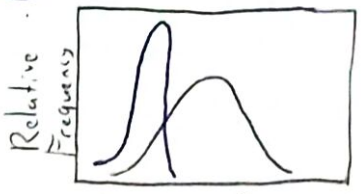
Two Identical Sets (Locations & Variabilities Same)

② 

Locations Differ

③ 

Variabilities Differ

④ 

Locations & Variabilities Differ

Covariance : A statistical measure which is defined as a systematic relationship between a pair of random variables wherein change in one variable responded by an equivalent change in another variable.

Range of Cov. $S_{xy} \sim (-\infty, +\infty)$

$Cov_{xy} = \emptyset$: Independent variables

$Cov_{xy} > \emptyset$: Two variables move together in same direction

$Cov_{xy} < \emptyset$: Two variable move together in opposite direction

(*) Doesn't show correlation directly. For example; if one variable doubles than covariance doubles!

Sample Covariance Formula : $S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Population Covariance Formula : $\sigma_{xy} = \sum_{i=1}^{n} \frac{(x_i - \mu_x)(y_i - \mu_y)}{N}$

Correlation (Coefficient) : Same as covariance, but in different range

Range : $r_{xy} \sim [-1, 1]$

$r = \emptyset$: Independent variables (no correlation)

$r = 1$: Perfect positive correlation

$r = -1$: Perfect negative correlation

Sample Correlation Formula : $r_{xy} = \frac{S_{xy}}{S_x S_y}$

Population Correlation Formula : $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

(*) r and ρ → greek letter "rho"

Covariance & Correlation Example :

⑤