EE2211 Pre-Tutorial 2

Dr Feng LIN
feng_lin@nus.edu.sg

Agenda

- Recap
- Self-learning
- Tutorial 2

Recap

Types of data

Data wrangling and cleaning

Data integrity and visualization

View Data by Scale/Level of Measurement

Nominal

- Lowest Level of Measurement
- Discrete Categories
- **NO** natural order
- Estimating a mean, median, or standard deviation, would be meaningless.
- Possible Measure: mode, frequency distribution

Ordinal

- **Ordered** Categories
- Relative Ranking
- Unknown "distance" between categories: orders matter but not the difference between values
- Possible Measure: mode, frequency distribution + median

View Data by Scale/Level of Measurement

Interval

- Ordered Categories
- Well-defined "**unit**" measurement:
- **Equal Interval**
- **Zero is arbitrary** (not absolute), in many cases human-defined
- **Ratio is meaningless**
- Possible Measure: mode, frequency distribution + median + mean, standard deviation, addition/subtraction

Ratio

- Most precise and **highest** level of measurement
- Ordered
- Equal Intervals
- **Natural Zeros**
- Possible Measure: mode, frequency distribution + median + mean, standard deviation, addition/subtraction + multiplication and division (ratio)

View Data by Levels/Scales of Measurement

Highest Precision

NOIR

Nominal

Named

Indicates difference

Ordinal

Named + Ordered

Indicates direction of difference

Interval

Named + Ordered + Equal Interval

Indicates amount of difference

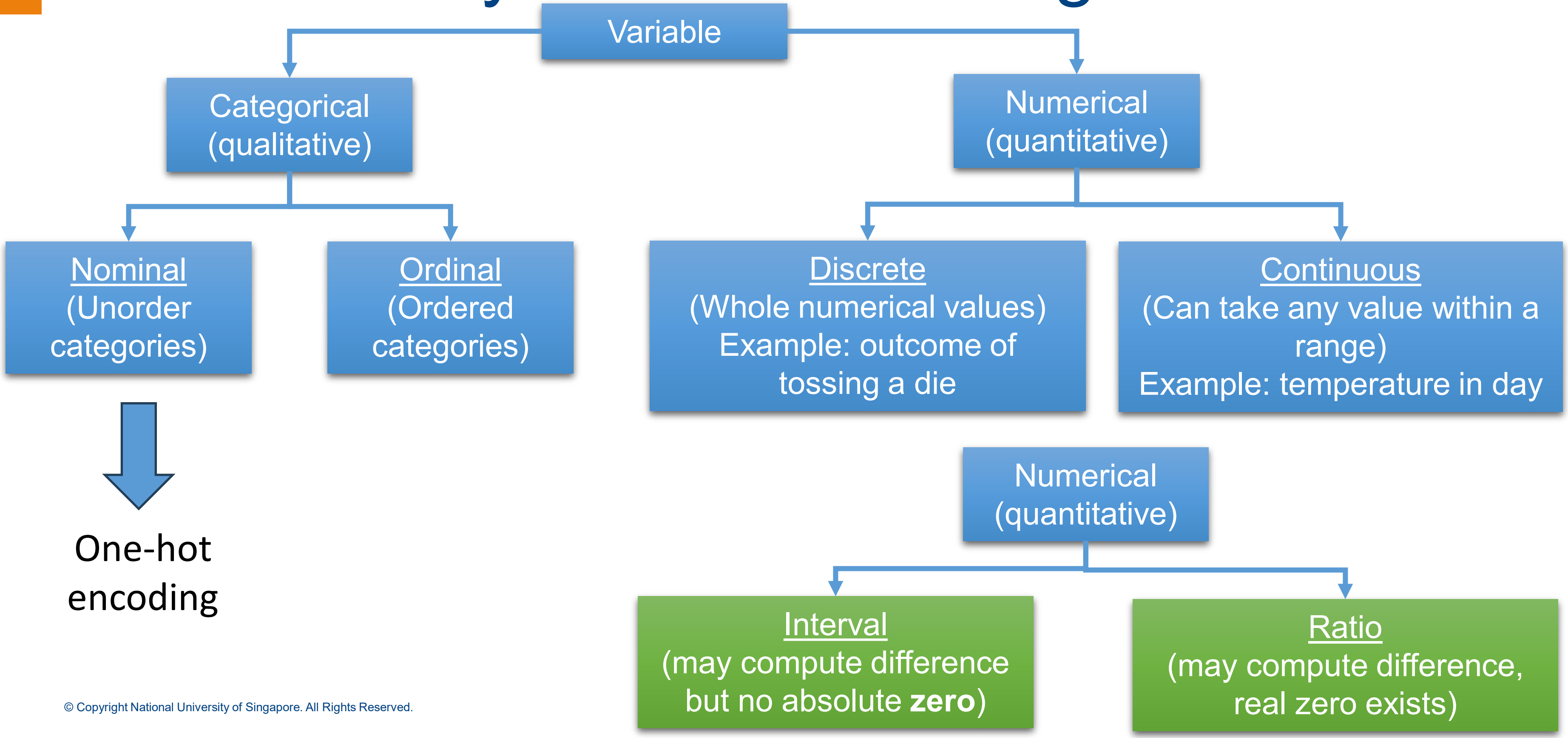
Ratio

Named + Ordered + Equal Interval + Has "True" Zero

"Absolute Zero"
"Absolute magnitudes"

Lowest

View Data by Numerical/Categorical



Data Wrangling

Determine your goal

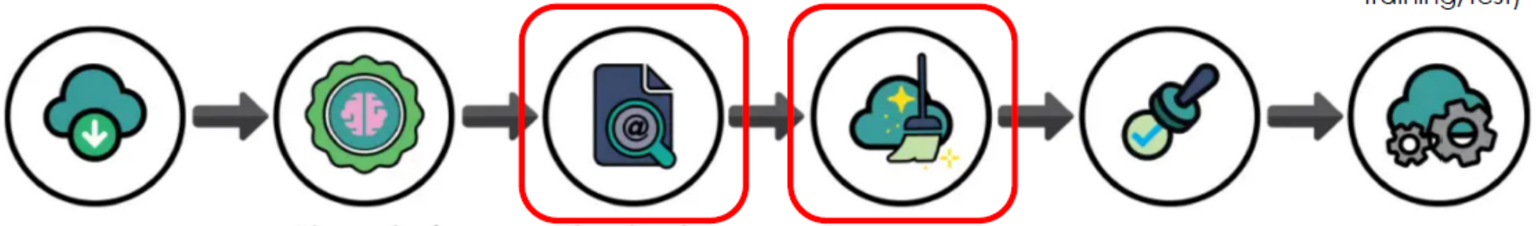
Make the most of dataset

Example: social network stored as graphs

Remove invalid data

Ensure data correctness

Use data (feature extraction/ training/test)



Extracting the Data

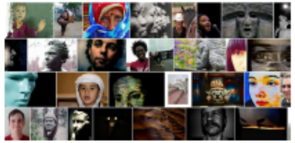
Discovering/ Analysing the Structure of the Data

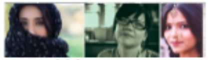
Choosing the Correct Format for the Data

Cleaning

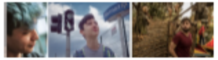
Validating

Deploying





Women



Men

Unifying format to .png

Remove noisy samples

Checking that all images have labels.

Use data for feature extraction and training your face detector!

Collect Human Face Images for Face Detector

Credit:https://understandingdata.com/what-is-data-wrangling/

Binary Coding

- One-hot encoding: unify several entities within one vector

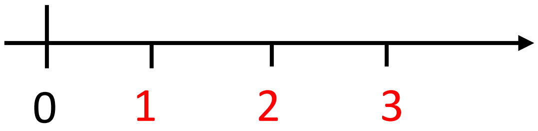
	Color		
	IsRed	**IsYellow**	**IsGeen**
Apple	1	0	0
Banana	0	1	0
Watermelon	0	0	1

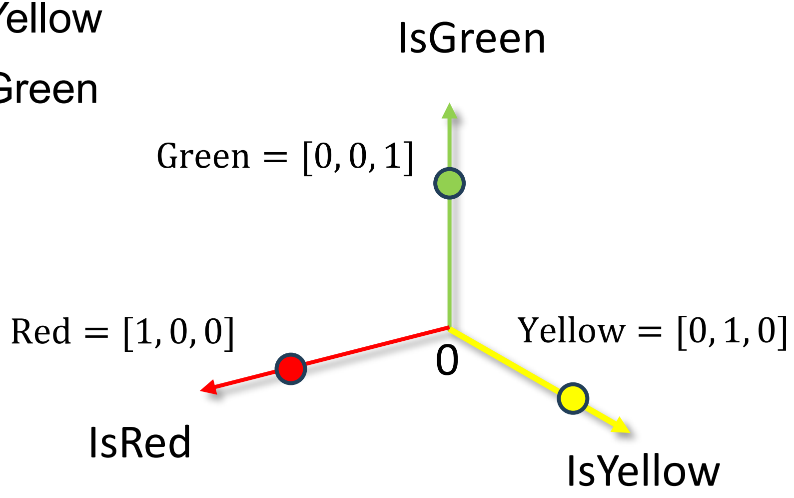
Red

Yellow

Green

$$\text{color} \in \{\overset{1}{\text{Red}}, \overset{2}{\text{Yellow}}, \overset{3}{\text{Green}}\}$$





Normalization

Often we have feature vectors in which features are on different scales.

For example:

$$\mathrm{x}_1 = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \qquad \ldots, \mathrm{x}_1 = \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}$$

First feature: Height $\in [140, 195]$

Second feature: Shoe size $\in [6, 13]$

- So even if both features are deemed equally “important”, unfortunately, any machine learning method would place more importance on the first feature because of its larger values, which is not ideal.
- Thus, we have to scale or normalize the features so that their dynamic ranges are roughly the same.

Normalization

- Min-max scaling

Define the minimum and maximum values of feature 1 to be

Max $$x_{max,1} = \max_{1 \le i \le n} x_{i1}$$

Min $$x_{min,1} = \min_{1 \le i \le n} x_{i1}$$

Then we create the normalized 1st features associated to each training sample as

$$\bar{x}_{i1} = \frac{x_{i1} - x_{min,1}}{x_{max,1} - x_{min,1}}$$

We can do this for all features so that, in some sense, they are all “normalized”.

Normalization

- Z-Score

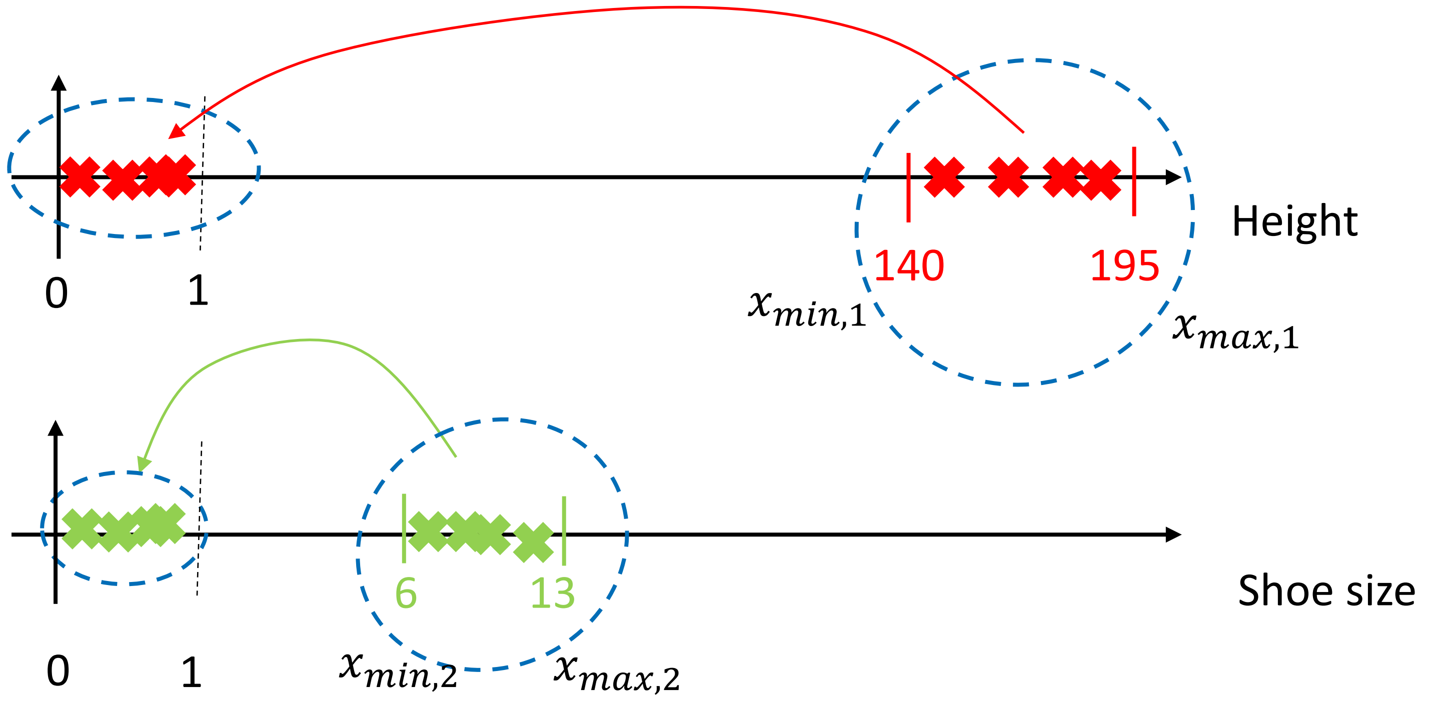
First we calculate the empirical mean and empirical standard deviation of each feature.

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1} \quad \text{and} \quad \sigma_1 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\mu_1)^2}$$

Then we create the normalized 1st features associated to each training sample as

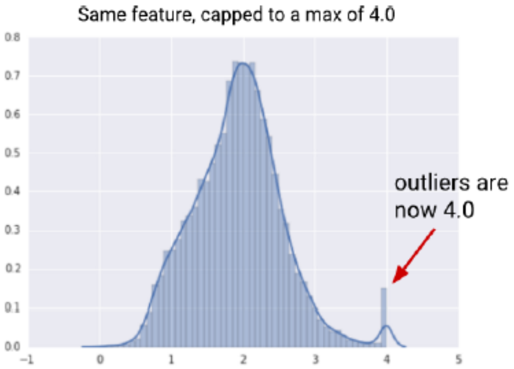
$$\bar{x}_{i1} = \frac{x_{i1}-\mu_1}{\sigma_1}$$

Normalization



Data Cleaning

- The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.
- Example:
  - Clipping outliers



  - Handling missing features

Students	Year of Birth	Gender	Height	GPA
Tan Ah Kow	1995	M	1.72	4.2
Ahmad Abdul	X NA	M	1.65	4.1
John Smith	1995	M	1.75	X NA
Chen Lulu	1995	F	X NA	4.0
Raj Kumar	1995	M	1.73	4.5
Li Xiuxiu	1994	F	1.70	3.8

Data Cleaning: Handling missing features

1. Removing the examples with missing features from the dataset
   – Can be done if the dataset is big enough so we can sacrifice some training examples

2. Using a learning algorithm that can deal with missing feature values
   – Example: random forest

3. Using a data imputation technique

Data Cleaning: Handling missing features: **Imputation**

- Method 1. Replace the missing value of a feature by an average value of this feature in the dataset:

$$\hat{x}^{(j)} \leftarrow \frac{1}{N} \sum_{i=1}^{N} x_i^{(j)}$$

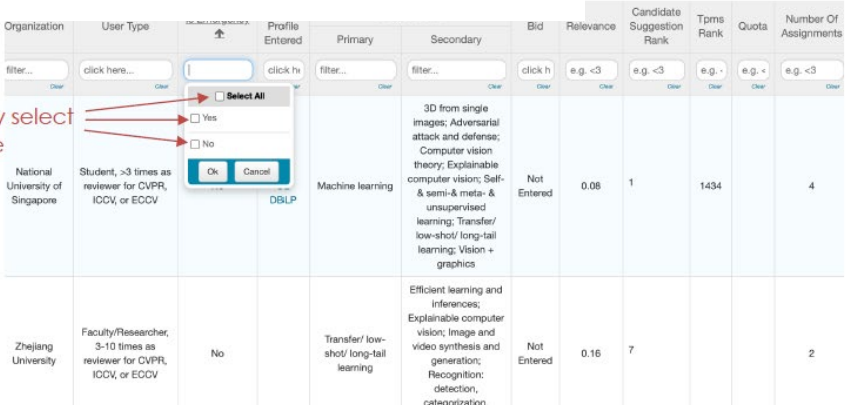
- Method 2. Highlight the missing value
  - Replace the missing value with a value outside the normal range of values.
  - For example, if the normal range is [0, 1], then you can set the missing value to −1.
  - Enforce the learning algorithm to learn what is best to do when the feature has a value significantly different from regular values.

Data Integrity

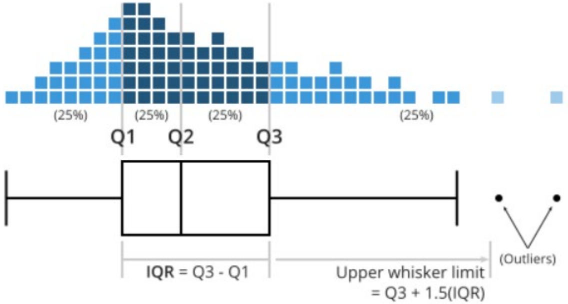
- Data integrity is the maintenance and the assurance of data accuracy and consistency;
  - A critical aspect to the <u>design</u>, implementation, and usage of any system that stores, processes, or retrieves data.
  - Very broad concept!
- Example:
  - In a dataset, numeric columns/cells should not accept alphabetic data.
  - A binary entry should only allow binary inputs

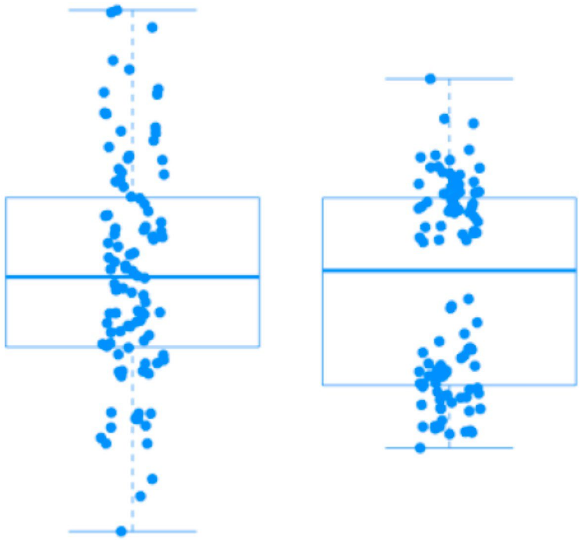
Mr. Mark John	33	21-08-1985	180	M	0433010010	Mel,VIC
Mr. Chris, Peter	34	21-Sep-1982	?	Fale	0000000000	Syd, NSW
Ethan Steedman	36	01/01/82	17o	M	0388886789	Mel,VIC

We can only select one of these



Visualization: Boxplots





Maximum (100th percentile) $Q_3 + 1.5 \times \mathrm{IQR}$

Third Quartile (75th percentile)

Median (50th percentile)

First Quartile (25th percentile)

Minimum (0th percentile) $Q_1 - 1.5 \times \mathrm{IQR}$

- The first quartile (Q_1) is defined as middle number between the smallest number and the median of the data set.
- The third quartile (Q_3) is defined as middle number between the highest number and the median of the data set.
- Interquartile range (IQR) is defined as distance between the first and third quartile, $\mathrm{IQR} = Q_3 - Q_1$.

Summary

- Types of data
  - NOIR
- Data wrangling and cleaning

- Data integrity and visualization
  - Integrity: Design
  - Visualization: Graphical Representation

THANK YOU