EE2211 Pre-Tutorial 2

Dr Feng LIN
feng_lin@nus.edu.sg

Agenda

- Recap
- Self-learning
- Tutorial 2

Recap

- Types of data
  - NOIR
  - Numerical/Categorical
- Data formatting
  - One-hot encoding
  - Normalization
- Visualization: Boxplots

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

Ratio

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

NOIR

Interval

Named + Ordered + Equal Interval

Ordinal

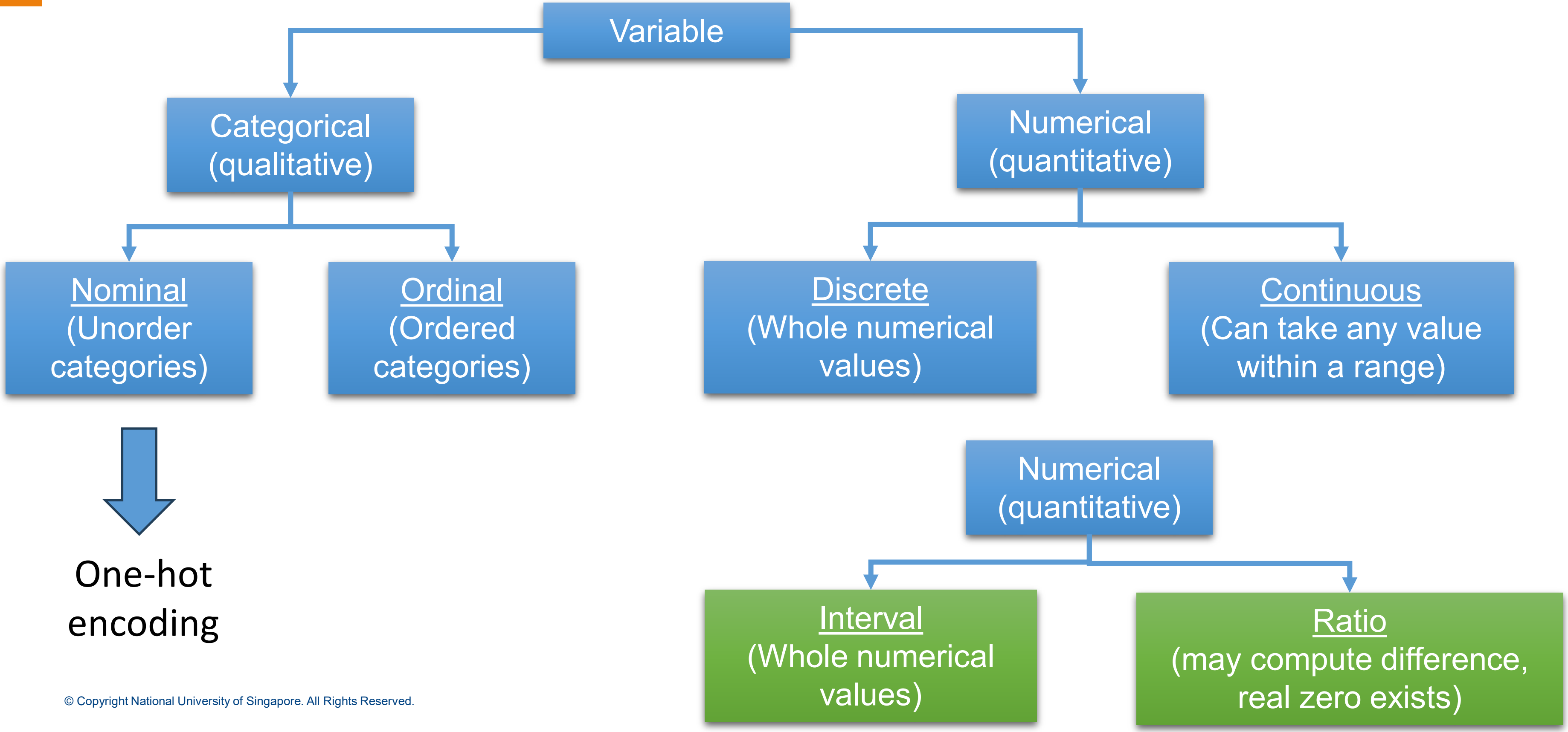
Named + Ordered

Nominal

Named

Lowest

View Data by Numerical/Categorical



Numerical Data

- Numerical data, as the name suggests, is data that represents numbers.

$$\mathrm{x}_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1d} \end{bmatrix}, \cdots, \mathrm{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nd} \end{bmatrix}$$

Feature Vector

Feature / attribute

Labels

$$D = \{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}$$

Numerical Data

- <u>Discrete Data</u>: These are data that are distinct and separate and can only take on certain values (usually finitely many values). This type of data can be counted.

For example, "the number of heads in 100 coin flips" is discrete numerical data because they can only take on values in the set of 101 Values $\{0,1,2, \dots, 99, 100\}$.

Numerical Data

- <u>Continuous Data</u>: These are data that cannot be counted but they can be measured.

  Example1: Temperature

  Example2: Height and weight of students

Categorical Data

Categorical data represent characteristics. There are two main types.

Nominal Data: These data represent discrete units and are used to represent variables that have no natural quantitative value. They are nothing but “labels”.

	Color
Apple	Red
Banana	Yellow
Watermelon	Green

$$\text{color} \in \{\text{Red}, \text{Yellow}, \text{Green}\}$$

Formatting Data: One-hot encoding

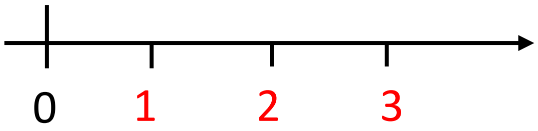
Categorical Data

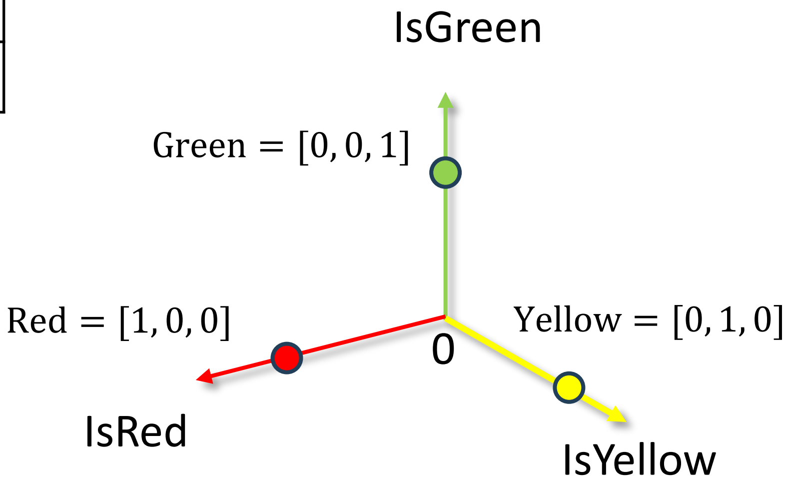
- One-hot encoding for nominal data

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

1 2 3

$$\text{color} \in \{\text{Red}, \text{Yellow}, \text{Green}\}$$





Categorical Data

Ordinal Data: These data represent discrete and ordered units. It is therefore nearly the same as nominal data, except that its ordering matters.

	Color			Quality
	IsRed	IsYellow	IsGeen	
Apple	1	0	0	Poor
Banana	0	1	0	Aaverage
Watermelon	0	0	1	Good

{Poor, Aaverage, Good, Excellent}

Ordered

1 2 3 4

Normalization

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

For example:

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

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

- 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

- Min-max scaling

Define the minimum and maximum values of feature 1 to be

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

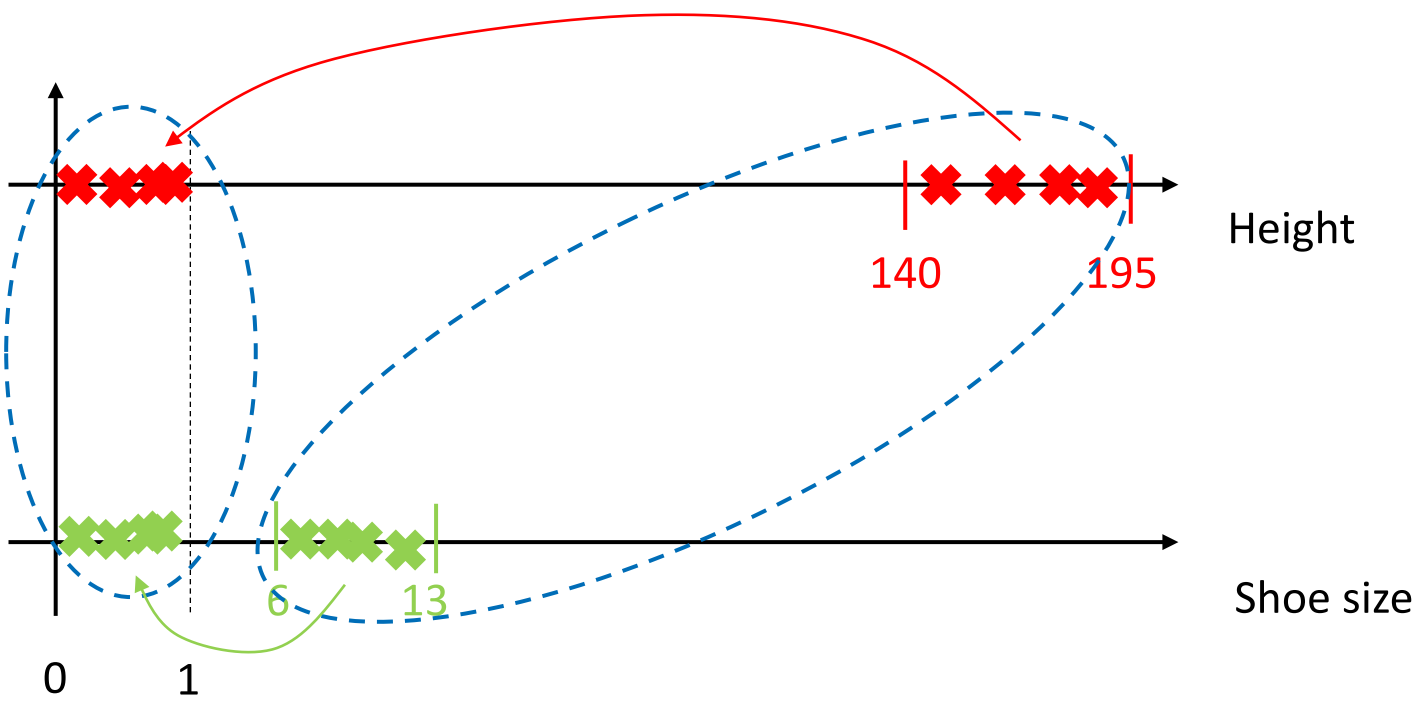
Min $$x_{min,1} = \min_{1 \le i \le n} x_{i1}$$

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

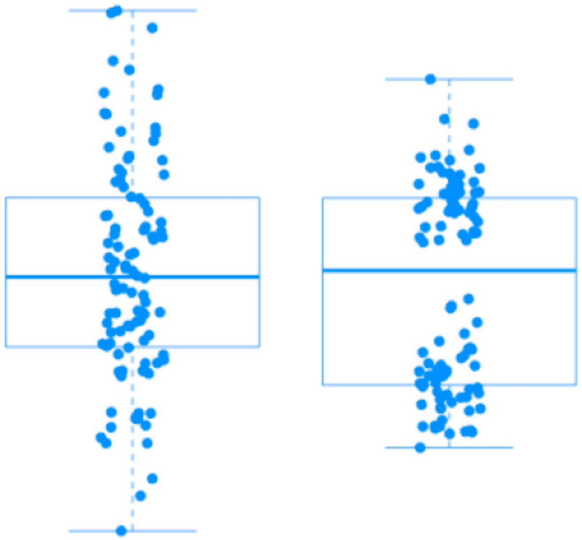
$$\bar{x}_{i1} = \frac{x_{i1} - x_{min,1}}{x_{max,1} - x_{min,1}}$$

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

Normalization



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, $\mathrm{IQR} = Q_3 - Q_1$

THANK YOU