THE UNIVERSITY OF MELBOURNE

Workshop 3

COMP20008

Elements of Data Processing

Zijie Xu

Agenda

- Slides repo
- Data pre-processing
- Data validation
- Data sampling

Workshop slides

- Slides from my workshops to be uploaded to GitHub repo: ccijjj/COMP20008-24s1
- Note: these slides are NOT a comprehensive summary of either lecture or workshop materials
- Please let me know if you find any errors

Data pre-processing

- Extracted data can be inconsistent
  - Different naming representations (UniMelb, UoM, Melbourne Uni)
  - Different formats (Age = 25.0, “Twenty-five”, “25yo”)
  - Clashes (Students with same student ID)
  - Outliers
- Pre-processing involves making the data consistent

Data pre-processing

- There could be missing data
  - MCAR: missing values unrelated other variables and itself
  - MAR: missing values related to other variables
  - MNAR: missing values related to the values of that variable itself
  - Disguised missing data: “unusual” or suspicious values
- Pre-processing includes imputing the missing values
  - Simple strategies for imputations: mode, mean, median

Data pre-processing

Course	Mark
B-Sci	89
B-Sci	
B-Sci	33
B-Sci	67
MC-DATASC	47
MC-DATASC	

Course	Mark
B-Sci	89
B-Sci	75
B-Sci	33
B-Sci	67
MC-DATASC	
MC-DATASC	

Course	Mark
B-Sci	89
B-Sci	75
B-Sci	
B-Sci	67
MC-DATASC	
MC-DATASC	92

Course	Mark
B-Sci	100
B-Sci	100
B-Sci	0
B-Sci	100
MC-DATASC	47
MC-DATASC	92

Data validation

- We validate data to ensure the quality of dataset
- Data validation involves checking for
  - Semantic errors: data inconsistent with intended meaning or purpose
  - Range errors: data falls outside the expected range of values
  - Format errors: data not in the expected format or structure
  - as well as other potential data quality issues such as missing or duplicate data

Data sampling

- We sample data to obtain a few examples that represents all data
  - with very large datasets, sampling can help reduce the computational resources required to process the data
- Some sampling methods:
  - Random sampling (with or without replacements)
  - Stratified sampling: Split population into groups depending on the relevant features, then sample from each group
- Samples should be representative, balanced, also large enough to be reasonably trustworthy

THE UNIVERSITY OF MELBOURNE

Thank you

More Resources: Canvas