VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

AIML427 Big Data

Data pre-processing

Dr Bach Hoai Nguyen

School of Engineering and Computer Science

Victoria University of Wellington

Bach.Hoai.Nguyen@ecs.vuw.ac.nz

Data Type

- Different types of data:
  - Continuous/real: 1.0, 1.05, 2.0, etc
  - Discrete:
    - Categorical/nominal: red, green, blue (ordered, distance)
    - Ordinal: Very happy > happy > OK (ordered, distance)
    - Integer: 4 > 2 > 1 (ordered, distance)
  - Other/special types of data (multi-media data):Text data, hyperlink data, image data

Data Pre-processing

- Normalisation/scaling:
  - adjust features with different scales to have the same scale
  - very important for distance-based algorithms such as KNN or SVM
  - Min-max normalisation: convert [X_{min}, X_{max}] to the range [0,1]

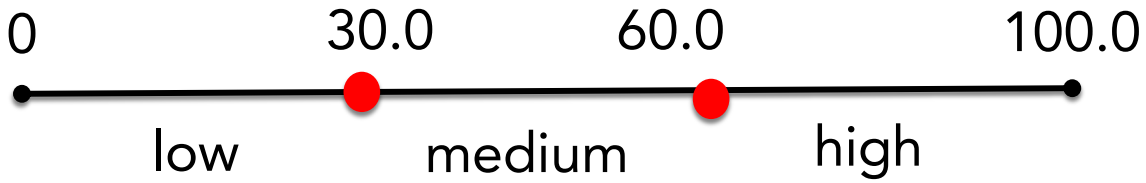
$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  - Standardisation: convert data to have a mean of 0 and standard variation of 1

$$X_{changed} = \frac{X - \mu}{\sigma}$$

Data Pre-processing

- Discretisation: convert a numeric attribute to a nominal attribute
  - e.g. Temperature attribute from {20.0, 50.0, 80.0} to {low, medium, high}



  - Unsupervised: does not consider the target output (class label in classification)
    - Equal-Width: each interval has the same width.
    - Equal-Depth: each interval has the same number of values.
  - Supervised: considers the target output
    - Entropy based method: repeatedly find splitting values to maximise information gain
    - one-rule decision tree algorithm (1RD)

Data Pre-processing

- Missing data
- Noisy data
- Outliers, unbalanced data
- Redundant data