Data Mining, CSCI 347, Fall 2019
Inputs to Data Mining, August 28

Some use the terms:

Data – recorded facts
Information – set of patterns, or expectations, that underlies the data
Knowledge – the accumulation of your set of expectations
Wisdom – value attached to knowledge

Data mining - techniques for finding and describing structural patterns in data to help explain that data and make predictions from it

Finding correlations, not causation

Use data mining techniques for two purposes:

1. Performance: given a new instance, predict what the outcome will be
   Black box such as instance based techniques
   Explanatory where see a structural pattern
2. Knowledge: learn a structural pattern

Attributes can be:

- nominal – discrete values
- numeric – continuous

Data mining algorithms typically work on individual, independent instances, with a predetermined set of attributes.

Useful terms for "levels of measurement": Nominal, Ordinal, Interval and Ratio

Nominal quantifies (in R nominal and ordinal values are called category variables or factors. Their values are levels.)

- Nominal comes from the Latin word for name
- Values serve only as labels, no order is assumed
- Operations: =, ≠

Ordinal quantifies

- Values are ordered, but no distance between values is defined
- Example
  - "temperature" from weather data has values "hot," "mild," and "cool". Could say hot > mild > cool
  - Can do comparisons, example "If temperature < hot then play=yes" (i.e. saying if temperature is mild or cool)
  - Sometimes distinction between nominal and ordinal is blurred (e.g. attribute outlook
- Operations: =, ≠, <, ≤, >, ≥

Interval quantifies

- Values are ordered and measured in fixed and equal units, but zero is not defined
- Can determine the distance between two values and determine the average
- Example - “temperature” expressed in degrees Fahrenheit
- Operations: =, ≠, <, ≤, >, ≥, + , -

Ratio quantifies

- Values are ordered, measured in fixed and equal units, and zero is defined
- All mathematical operations are allowed
- Example - “temperature” expressed in degrees Kelvin since in the Kelvin scale 0 is absolute 0.
- Operations: =, ≠, <, ≤, >, ≥, + , -, *, /

Metadata – data about the data

Data mining rarely uses metadata, beyond attribute names and types

Information from text and Nguyen Hung Son,
http://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf

Data in the real world is dirty

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- noisy: containing errors or outliers
- inconsistent: containing discrepancies in codes or names

Data cleaning - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

- Fill in missing values, use some set value or maybe the mean
- Converting nominal to numeric
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Data integration

- Integration of multiple datasets, databases, files, etc.
    - Example - have city and state, want counties

Missing data

- May be due to:
    - equipment malfunction
    - inconsistent with other recorded data and thus deleted
    - data not entered due to misunderstanding
    - certain data may not be considered important at the time of entry