TIM 245 Lecture 2 (4/5/17)

Agenda

1) Review of Lecture 1

2) Data

3) Data preprocessing

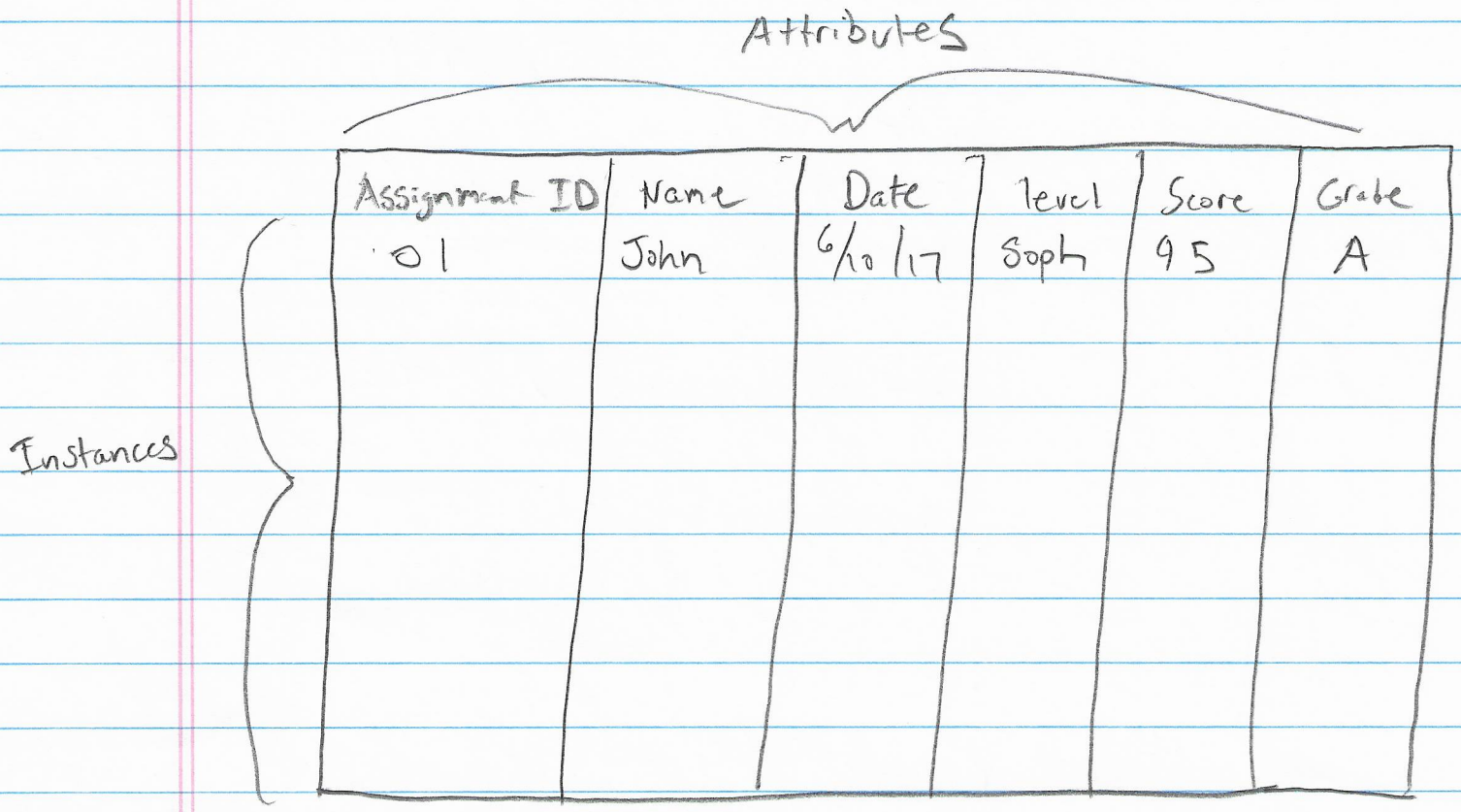
4) Project Kick off and Project Phase I assignment

5) work on project

② Data

Dataset: collection of **instances** (rows) across a set of **attributes** (columns)

Example: class grades



Types of attributes

Nominal = Value are distinct ($=, \neq$)

Example: name

Ordinal: Values are distinct and can be ordered ($=, \neq, \leq, >$)

Example: Level

Interval: Values are distinct, can be ordered, and have a meaningful difference ($=, \neq, <, >, +, -$)

Ratio: value are distinct, can be ordered, and have a meaningful difference and ratio ($=, \neq, <, >, +, -, *, /$)

Example: score

Types of data-sets

Record Data (Tables) : flat files (csv)
hierarchical files (XML, JSON)
Relational databases (SQL)

Time series Data (Ordered data) : flat files (csv)

Text Data : hierarchical files (XML, JSON)

Graph Data : hierarchical files (XML, JSON)
Graph databases (Neo4J, MongoDB)

Where to get data:

Data mining Competitions	:	Kaggle Data Driven
Website API	:	Facebook Twitter Google Yelp Zillow
Website Scraping	:	Forum product listings
open Source Datasets	:	UCI Machine Learning Data.gov AWS Public Datasets
Create your own data	:	Wearable devices, research experiments

③ Data Preprocessing

Why do we need to do pre-processing?

Real-World data is often

1) Incomplete: lacks certain attributes of interest

2) Noisy: contains errors or outlier values

3) Inconsistent: uses different values for the same thing (lingo problem)

Addressing these issues will improve the quality of the data mining results:

Descriptive Analysis: patterns that are more useful and relevant to the problem under consideration

Predictive Analysis: models that have better predictive accuracy and generalization

There are four basic steps in data preprocessing:

1) Data Cleaning: filling in missing values, smoothy noisy data removing outliers, removing duplicates
(Instances / Rows)

2) Data Integration: joining multiple datasets entity recognition repeat data cleaning
(attributes / columns)

3) Data Transformation: normalization binning
(instances / rows)

4) Data Reduction: attribute selection Dimensionality Reduction Sampling
(attributes / columns, Instances / rows)