University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2020
Lecture 04 (Sep 3)

Announcement

- Homework 1
  - posted in Canvas, due at 9:30am, Th, Sep 10
  - HW1 for CSCI 4502 vs. CSCI 5502
  - Jupyter Notebook for Q3
  - SUBMIT your attempt in Canvas before deadline
  - check syllabus for office hours

Jupyter Notebook

- CS Jupyter Hub
  - https://coding.csel.io/hub/login
  - default coding environment
- https://csel.cs.colorado.edu/coding_environment_jupyter_notebook.html
  - tutorial video

Review

- Chapter 2: Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Nominal Attributes

- E.g., courses taken by different students
- Method 1: simple matching
  - d(i, j) = (p - m) / p
  - m: # of matches, p: total # of variables
- Method 2: view each state as a binary variable
  - e.g., courses (C1, C2, C3, C4); then (0, 1, 0, 0) means C2

Binary Variables

		B	B	
		1	0	sum
A	1	q	r	q+r
A	0	s	t	s+t
	sum	q+s	r+t	q+r+s+t

- Contingency table
- Symmetric binary variables

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Asymmetric binary variables
  - Jaccard coefficient

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j)$$

Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender: symmetric
- Others: asymmetric
- Consider only asymmetric binary variables
- Y (yes) and P (positive) is 1, and N is 0
  - d(Jack, Mary) = (0+1)/(2+0+1) = 0.33
  - d(Jack, Jim) = (1+1)/(1+1+1) = 0.67
  - d(Jim, Mary) = (1+2)/(1+1+2)=0.75

$$d(i,j) = \frac{r+s}{q+r+s}$$

Ordinal Variables

- E.g., gold, silver, bronze
- Order is important: rank
- Treat like interval-scaled variables
  - map to their ranks
  - map to range [0, 1]

$$r_{if} \in \{1, \ldots, M_f\}$$

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- (1, 2, 3) => (0.0, 0.5, 1.0)
- dissimilarity of interval-scaled variables

Cosine Similarity Example

- D1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
- D2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)
- D1• D2 = 5x3 + 0x0 + … + 0x1 = 25
- ||D1|| = (5x5 + 0x0 + … + 0x0)^(1/2) = 6.481
- ||D2|| = 4.12, cos(D1, D2) = 0.936

$$s(x, y) = \frac{x^t \cdot y}{||x|| ||y||}$$

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Variables of Mixed Types

- Data may contain different types of variables
- Weighted combination

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

- $\delta_{ij}^{(f)} = 0$ if
  - x_{if} or x_{jf} is missing
  - $x_{if} = x_{jf} = 0$ and f is an asymmetric binary variable
- otherwise = 1

Summary

- Chapter 2: Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Discussion

- Given a dataset: e.g., Twitter, Sports, News, Traffic, …
- What attribute types you may be able to use?
- What knowledge you may be able to learn?
- How would that knowledge be useful?
- …

University of Colorado
Boulder

Chapter 3:
Data Preprocessing

Chapter 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization

Measures of Data Quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability
- Accessibility

Major Tasks in Preprocessing

- Data cleaning
  - fill in missing values, smooth noisy data, identify or remove outliers, resolve inconsistencies
- Data integration
  - integration of multiple data sources
- Data reduction
  - dimensionality, numerosity, compression
- Data transformation and data discretization
  - normalization, concept hierarchy generation

Chapter 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization

Why Data Cleaning?

- Imperfect real-world data
- Incomplete: missing attributes, values
  - e.g., age = "", major = ""
- Noisy: containing errors or outliers
  - e.g., salary = "-10"
- Inconsistent: containing discrepancies
  - e.g., age = "21", birthday = "08/03/1995"
  - e.g., ratings of "1, 2, 3" and "A, B, C"

Why Are Data Imperfect?

- Incomplete data
  - "not applicable" values
  - time between collection and analysis
  - human/hardware/ software problems
- Noisy data
  - faulty data collection instruments
  - human or computer error at data entry
  - errors in data transmission

Why Are Data Imperfect?

- Inconsistent data
  - different data sources
  - naming conventions, data formats
    - e.g., date "03/07/11"
  - functional dependency violation
    - e.g., modify some linked data
- No quality data, no quality data mining results!

How to Handle Missing Data?

- Ignore the tuple
- Fill in the missing value manually
- Fill in it automatically with
  - global constant; attribute mean; attribute mean of the same class
  - most probable value: e.g., regression, Bayesian inference, decision tree

How to Handle Noisy Data?

- Binning
  - first sort & partition data into bins
  - then smooth by
    - bin means
    - bin median
    - bin boundaries

Sorted data for *price* (in dollars): 4, 8, 15,

Partition into (equal-frequency) bir

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

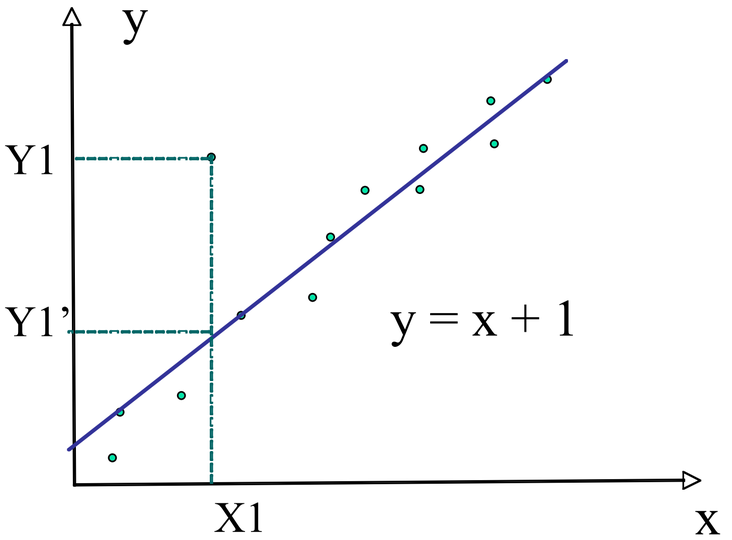
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

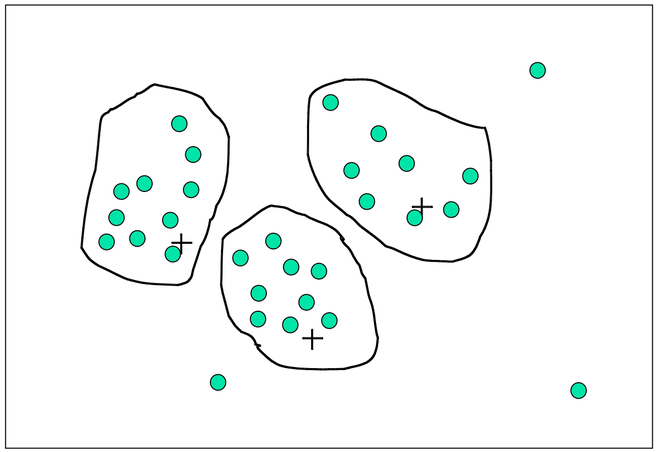
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

How to Handle Noisy Data?

- Regression
  - fit data into regression functions



- Clustering
  - detect and remove outliers



Chapter 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization

Data Integration

- Combines data from multiple sources
- Entity identification
  - schema integration, object matching
  - e.g., student_id vs. student_number
- Redundant data
  - different naming, derived data
  - may be detected by correlation analysis