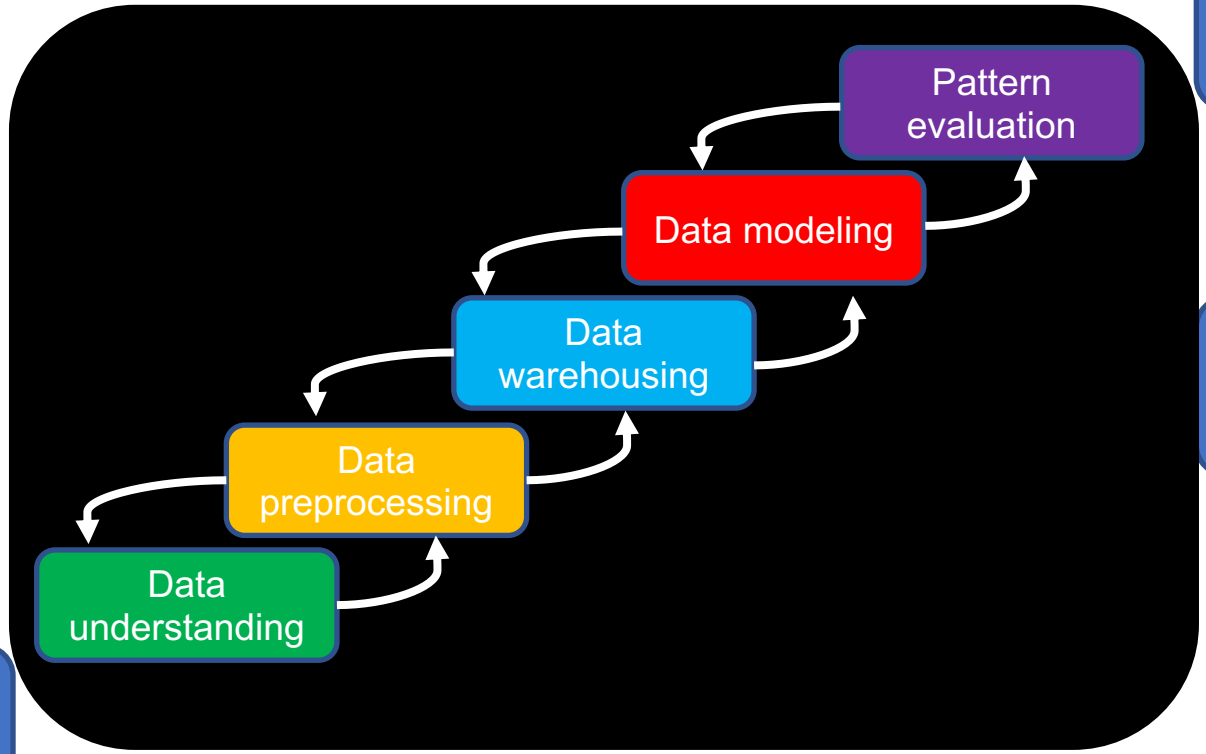
Data Understanding

Data Mining:
Data Mining Pipeline
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Describe the key properties of data. Apply techniques to characterize different datasets.

Data Mining Pipeline



Application

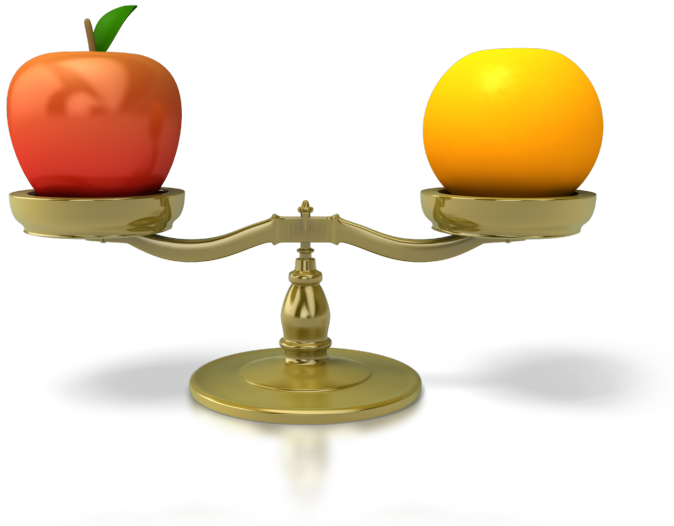
Knowledge

Technique

Data

Data Understanding

- Data objects & attributes
- Data statistics
- Data visualization
- Data similarity



Object Similarity/Dissimilarity

➢ Object matrix

- n objects x p attributes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

➢ Dissimilarity matrix

- n objects x n objects

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Nominal Attributes

	Car 1	Car 2
Type	SUV	Sedan
Make	Honda	Honda
Model	CR-V	Civic
Color	White	Silver
Condition	Good	Good

- ➢ Similarity
  - s = 1 if x = y; otherwise s = 0
- ➢ Dissimilarity
  - d = 0 if x = y; otherwise d = 1
- ➢ Customized
  - E.g., color: white is more similar to silver than red

Binary Attributes

- ➢ Symmetric (i.e., equal chance of Y or N)
  - Same calculation as nominal attributes
  - E.g., Hamming distance: #bits that are different
- ➢ Asymmetric (e.g., Y is less likely than N)
  - E.g., items purchased or medical symptoms

	Item 1	Item 2	Item 3	Gender	Cough	Fever
User A	Y	Y	N	F	N	N
User B	Y	N	N	M	Y	N

Binary Attributes

	B = Y	B = N	sum
A = Y	q	r	q+r
A = N	s	t	s+t
sum	q+s	r+t	q+r+s+t

- Symmetric variables

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

- Asymmetric variables

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

- Jaccard coefficient

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

Ordinal Attributes

- E.g., Gold, Silver, Bronze
- Map to their ranks $r_{if} \in \{1, \ldots, M_f\}$
  - => (1, 2, 3)
- Map to [0, 1] $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
  - => (0, 0.5, 1.0)
- Dissimilarity between the mapped values

Numeric Object Dissimilarity

- Usually measured by distance
- Minkowski distance (l_p norm)

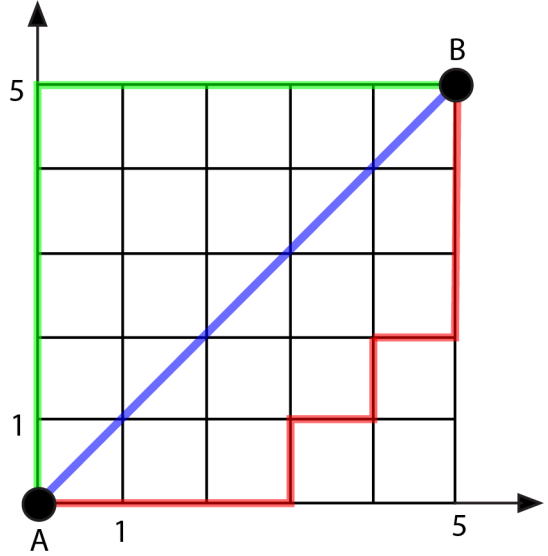
$$d(i,j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p}$$

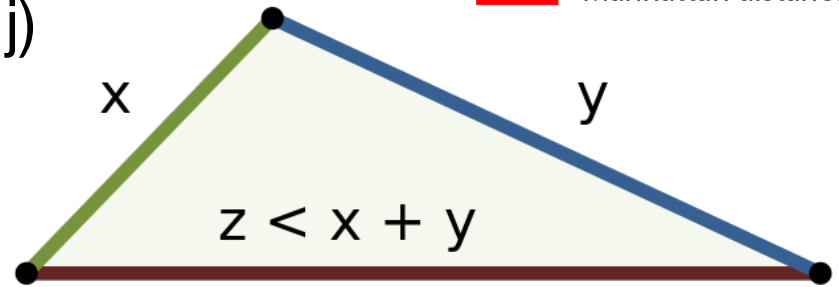
- p = 1: Manhattan distance
- p = 2: Euclidean distance

Distance Measure

➢ Properties

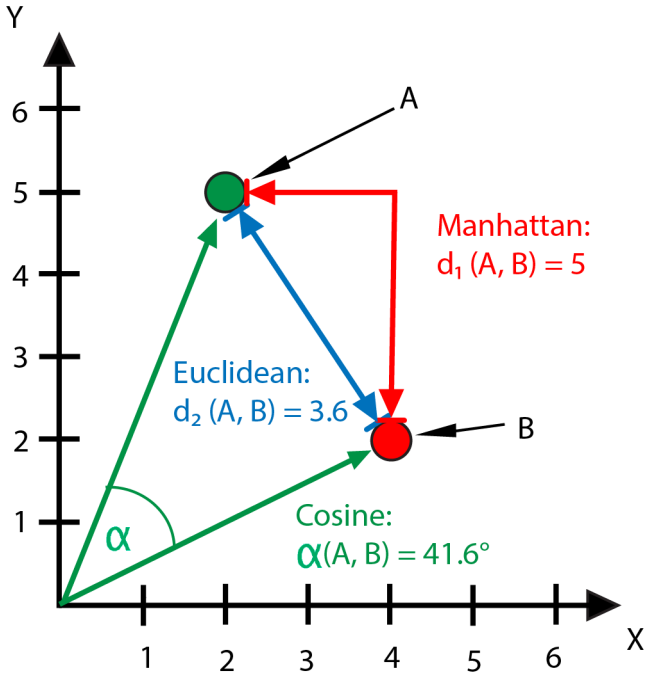
- d(i,i) = 0
- d(i,j) >= 0
- d(i,j) = d(j,i)
- d(i,j) <= d(i, k) + d(k, j)
- triangular inequality





Cosine Similarity

- E.g., text documents
  - Frequency of word occurrence
  - High dimensional, sparse
- Angular similarity of vectors: inner product



$$cos(\theta) = cos(A, B) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}}$$

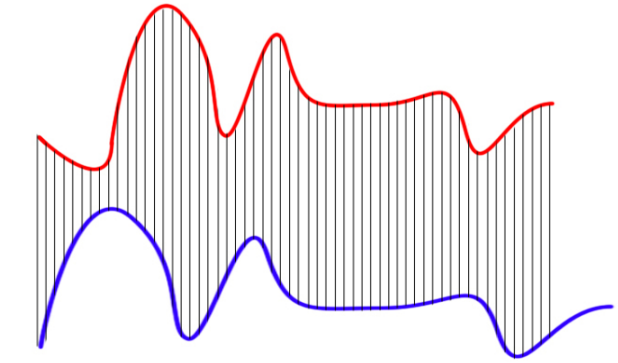
Cosine Similarity Example

- D1• D2 = 3x1+5x3+2x4+1x2+2x1 = 30
- ||D1|| = (3x3+5x5+2x2+1x1+2x2)^(1/2) = 6.557
- ||D2|| = (1x1+3x3+4x4+2x2+1x1)^(1/2) = 5.568
- cos(D1, D2) = 30 / (6.557 x 5.568) = 0.822

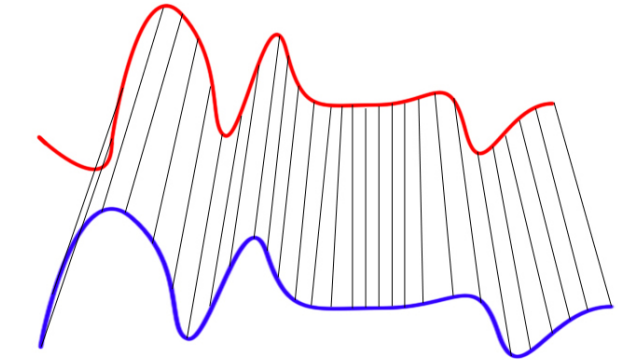
	game	basketball	player	injury	win
Document 1	3	5	2	1	2
Document 2	1	3	4	2	1

Sequential Data, Time Series

- Euclidean distance
- Dynamic time warping
- Minimum jump cost
- …



Euclidean Matching



Dynamic Time Warping Matching

Mixed Attribute Types

- Weighted sum across attributes

- Attribute selection, correlated attributes, normalization, weights, missing values, …

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

Data Similarity/Dissimilarity

➢ How to choose the right measure?

- Dense, continuous data: Euclidean, Manhattan, …
- Asymmetric attributes: ignore the null/null cases
- Sparse data: cosine similarity, Jaccard similarity
- Subset: e.g., seasonal patterns, subgroups
- Domain knowledge, types of pattern to learn

Case Study: Online Social Media

- Types of data
  - User, post, timestamp, reactions, network
- Statistics, visualization
  - #objects, #attributes, distributions, outliers
- Similarity/dissimilarity
  - Nominal, binary, numeric, text, temporal

Data Understanding

- Data objects & attributes
- Data statistics
- Data visualization
- Data similarity