Distance or Similarity Measures

Distance or Similarity Measures

- **Many data mining and analytics tasks involve the comparison of objects and determining in terms of their similarities (or dissimilarities)**
  - Clustering
  - Nearest-neighbor search, classification, and prediction
  - Characterization and discrimination
  - Automatic categorization
  - Correlation analysis
- **Many of todays real-world applications rely on the computation similarities or distances among objects**
  - Personalization
  - Recommender systems
  - Document categorization
  - Information retrieval
  - Target marketing

Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity (e.g., distance)**
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity refers to a similarity or dissimilarity**

Distance or Similarity Measures

- **Measuring Distance**
  - In order to group similar items, we need a way to measure the distance between objects (e.g., records)
  - Often requires the representation of objects as "feature vectors"

An Employee DB

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

Feature vector corresponding to Employee 2: <M, 51, 64000.0>

Term Frequencies for Documents

	T1	T2	T3	T4	T5	T6
Doc1	0	4	0	0	0	2
Doc2	3	1	4	3	1	2
Doc3	3	0	0	0	3	0
Doc4	0	1	0	3	0	0
Doc5	2	2	2	3	1	4

Feature vector corresponding to Document 4: <0, 1, 0, 3, 0, 0>

Distance or Similarity Measures

- **Properties of Distance Measures:**
  - for all objects A and B, $dist(A, B) \geq 0$, and $dist(A, B) = dist(B, A)$
  - for any object A, $dist(A, A) = 0$
  - $dist(A, C) \leq dist(A, B) + dist(B, C)$

- **Representation of objects as vectors:**
  - Each data object (item) can be viewed as an n-dimensional vector, where the dimensions are the attributes (features) in the data
  - Example (employee DB): Emp. ID 2 = <M, 51, 64000>
  - Example (Documents): DOC2 = <3, 1, 4, 3, 1, 2>
  - The vector representation allows us to compute distance or similarity between pairs of items using standard vector operations, e.g.,
    - Cosine of the angle between vectors
    - Manhattan distance
    - Euclidean distance
    - Hamming Distance

Data Matrix and Distance Matrix

- **Data matrix**
  - Conceptual representation of a table
    - Cols = features; rows = data objects
  - *n* data points with *p* dimensions
  - Each row in the matrix is the vector representation of a data object

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

- **Distance (or Similarity) Matrix**
  - *n* data points, but indicates only the pairwise distance (or similarity)
  - A triangular matrix
  - Symmetric

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

Proximity Measure for Nominal Attributes

- **If object attributes are all nominal (categorical), then proximity measure are used to compare objects**

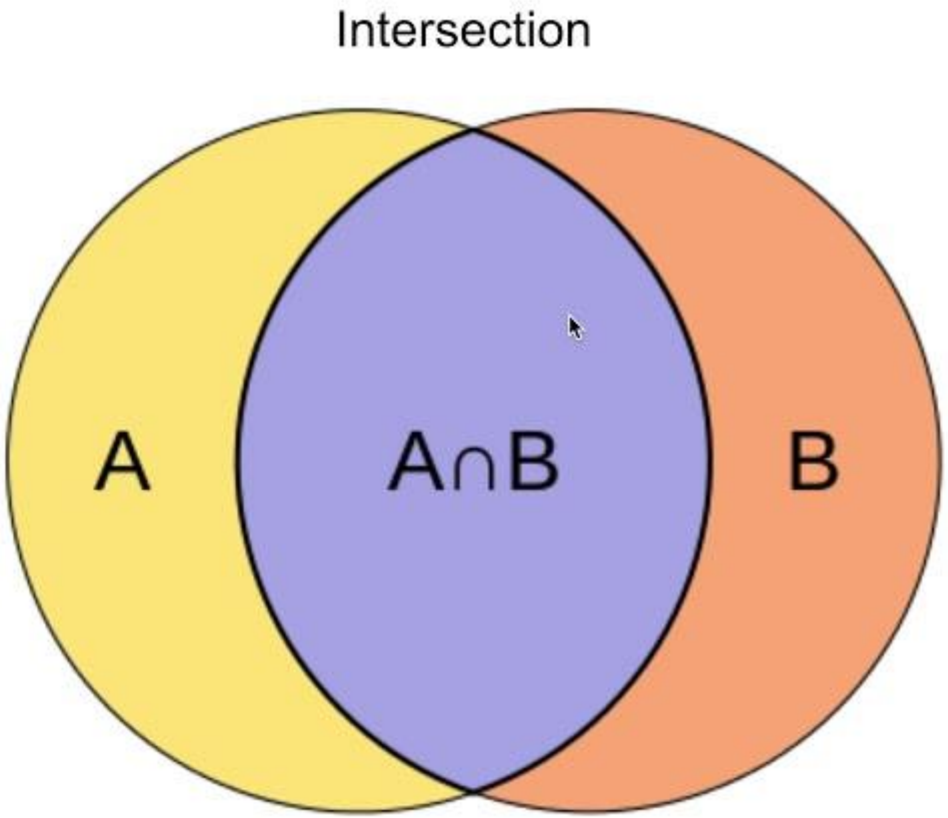
- **Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)**

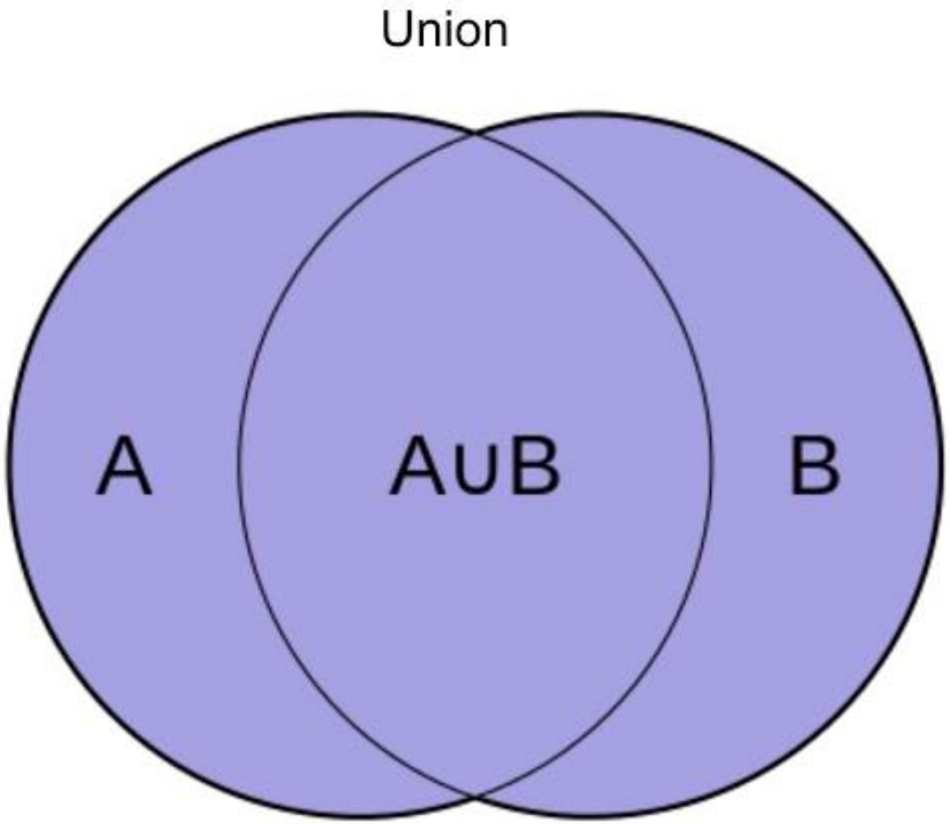
- **Method 1: Simple matching**
  - *m*: # of matches, *p*: total # of variables

$$\boxed{d(i,j) = \frac{p-m}{p}}$$

- **Method 2: Convert to Standard Spreadsheet format**
  - For each attribute *A* create *M* binary attribute for the *M* nominal states of *A*
  - Then use standard vector-based similarity or distance metrics

Jaccard coefficient





$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Normalizing or Standardizing Numeric Data

- **Z-score:**
  - x: raw value to be standardized, μ: mean of the population, σ: standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the value is below the mean, "+" when above

$$z = \frac{x - \mu}{\sigma}$$

- **Min-Max Normalization**

$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}(new\max - new\min) + new\min$$

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

ID	Gender	Age	Salary
1	1	0.00	0.00
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32

Common Distance Measures for Numeric Data

- **Consider two vectors**
  - Rows in the data matrix

$$\boxed{X = \langle x_1, x_2, \cdots, x_n \rangle} \quad \boxed{Y = \langle y_1, y_2, \cdots, y_n \rangle}$$

- **Common Distance Measures:**
  - Manhattan distance:

$$\boxed{dist(X,Y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|}$$

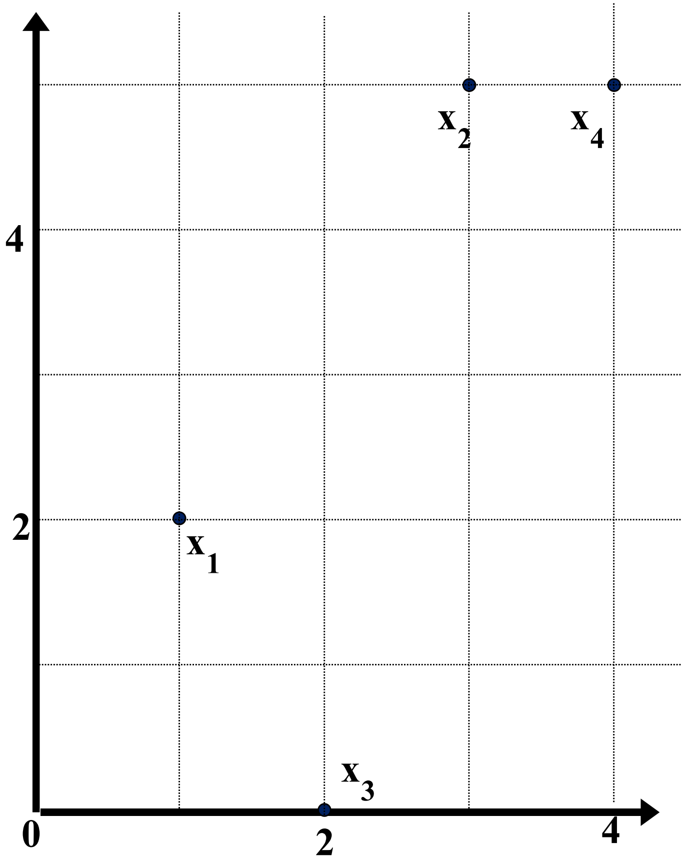
  - Euclidean distance:

$$\boxed{dist(X,Y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}}$$

  - Distance can be defined as a dual of a similarity measure

$$\boxed{dist(X,Y) = 1 - sim(X,Y)}$$

Example: Data Matrix and Distance Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Distance Matrix (Manhattan)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Distance Matrix (Euclidean)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

- **Minkowski distance: A popular distance measure**

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

  - where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

- **Note that Euclidean and Manhattan distances are special cases**
  - $h = 1$: (L_1 norm) Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

  - $h = 2$: (L_2 norm) Euclidean distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

Vector-Based Similarity Measures

- **In some situations, distance measures provide a skewed view of data**
  - E.g., when the data is very sparse and 0's in the vectors are not significant
  - In such cases, typically vector-based similarity measures are used
  - Most common measure: Cosine similarity

$$X = \langle x_1, x_2, \cdots, x_n \rangle \qquad Y = \langle y_1, y_2, \cdots, y_n \rangle$$

  - Dot product of two vectors: $sim(X,Y) = X \bullet Y = \sum_i x_i \times y_i$

  - Cosine Similarity = normalized dot product

  - the norm of a vector X is: $\|X\| = \sqrt{\sum_i x_i^2}$

  - the cosine similarity is:

$$sim(X,Y) = \frac{X \bullet Y}{\|X\| \times \|y\|} = \frac{\sum_i (x_i \times y_i)}{\sqrt{\sum_i x_i^2} \times \sqrt{\sum_i y_i^2}}$$

Example Application: Information Retrieval

- **Documents are represented as "bags of words"**
- **Represented as vectors when used computationally**
  - A vector is an array of floating point (or binary in case of bit maps)
  - Has direction and magnitude
  - Each vector has a place for every term in collection (most are sparse)

Document Ids

a document vector

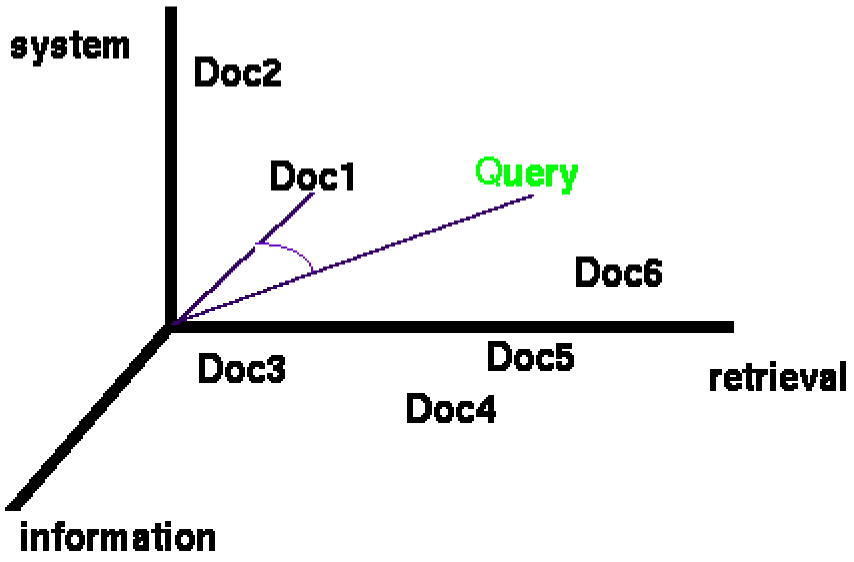
	nova	galaxy	heat	actor	film	role
A	1.0	0.5	0.3			
B	0.5	1.0				
C		1.0	0.8	0.7		
D		0.9	1.0	0.5		
E				1.0		1.0
F					0.7	
G	0.5		0.7			0.9
H		0.6		1.0	0.3	0.2
I			0.7	0.5		0.3

$$D_i = w_{d_{i1}}, w_{d_{i2}}, \ldots, w_{d_{it}}$$

$$Q = w_{q1}, w_{q2}, \ldots, w_{qt}$$

$$w = 0 \text{ if a term is absent}$$

Documents & Query in n-dimensional Space



- **Documents are represented as vectors in the term space**
  - Typically values in each dimension correspond to the frequency of the corresponding term in the document
- **Queries represented as vectors in the same vector-space**
- **Cosine similarity between the query and documents is often used to rank retrieved documents**

Example: Similarities among Documents

- **Consider the following document-term matrix**

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	0	4	0	0	0	2	1	3
Doc2	3	1	4	3	1	2	0	1
Doc3	3	0	0	0	3	0	3	0
Doc4	0	1	0	3	0	0	2	0
Doc5	2	2	2	3	1	4	0	2

Dot-Product(Doc2,Doc4) = <3,1,4,3,1,2,0,1> * <0,1,0,3,0,0,2,0>
0 + 1 + 0 + 9 + 0 + 0 + 0 + 0 = 10

Norm (Doc2) = SQRT(9+1+16+9+1+4+0+1) = 6.4
Norm (Doc4) = SQRT(0+1+0+9+0+0+4+0) = 3.74

Cosine(Doc2, Doc4) = 10 / (6.4 * 3.74) = 0.42

Correlation as Similarity

- In cases where there could be high mean variance across data objects (e.g., movie ratings), Pearson Correlation coefficient is the best option
- Pearson Correlation

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$corr(x,y) = \frac{\mathrm{cov}(x,y)}{stdev(x) \cdot stdev(y)}$$

- Often used in recommender systems based on Collaborative Filtering

*K*NN and Collaborative Filtering

- **Collaborative Filtering Example**
  - A movie rating system
  - Ratings scale: 1 = “hate it”; 7 = “love it”
  - Historical DB of users includes ratings of movies by Sally, Bob, Chris, and Lynn
  - Karen is a new user who has rated 3 movies, but has not yet seen “Independence Day”; should we recommend it to her?

	Sally	Bob	Chris	Lynn	Karen
Star Wars	7	7	3	4	7
Jurassic Park	6	4	7	4	4
Terminator II	3	4	7	6	3
Independence Day	7	6	2	2	?

Will Karen like “Independence Day?”

Collaborative Filtering

(*k* Nearest Neighbor Example)

	Star Wars	Jurassic Park	Terminator 2	Indep. Day	*Average*	Cosine	Distance	Euclid	Pearson
Sally	7	6	3	7	5.33	0.983	2	2.00	0.85
Bob	7	4	4	6	5.00	0.995	1	1.00	0.97
Chris	3	7	7	2	5.67	0.787	11	6.40	-0.97
Lynn	4	4	6	2	4.67	0.874	6	4.24	-0.69
Karen	7	4	3	**?**	4.67	1.000	0	0.00	1.00

K	Prediction
1	6
2	6.5
3	5

K is the number of nearest neighbors used in to find the average predicted ratings of Karen on Indep. Day.

Example computation:

$$\text{Pearson(Sally, Karen)} = \frac{(7-5.33)*(7-4.67) + (6-5.33)*(4-4.67) + (3-5.33)*(3-4.67)}{\text{SQRT}\left(((7-5.33)^2 + (6-5.33)^2 + (3-5.33)^2) * ((7-4.67)^2 + (4-4.67)^2 + (3-4.67)^2)\right)} = 0.85$$