Measuring distance/ similarity of data objects

Multiple data types

- Records of users
- Graphs
- Images
- Videos
- Text (webpages, books)
- Strings (DNA sequences)
- Timeseries
- **How do we compare them?**

Feature space representation

- Usually data objects consist of a set of attributes (also known as **dimensions**)
- J. Smith, 20, 200K
- If all d dimensions are **real-valued** then we can **visualize** each data point as points in a **d-dimensional space**
- If all d dimensions are **binary** then we can think of each data point as a **binary vector**

Distance functions

- The distance $d(x, y)$ between two objects $x$and y is a metric if
  - $d(i, j) \geq 0$ (non-negativity)
  - $d(i, i) = 0$ (isolation)
  - $d(i, j) = d(j, i)$ (symmetry)
  - $d(i, j) \leq d(i, h) + d(h, j)$ (triangular inequality)
- The definitions of distance functions are usually different for real, boolean, categorical, and ordinal variables.
- Weights may be associated with different variables based on applications and data semantics.

Data Structures

- data matrix

$$\text{tuples/objects}\left\{\overbrace{\begin{bmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{il} & \cdots & x_{id} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nl} & \cdots & x_{nd} \end{bmatrix}}^{\text{attributes/dimensions}}\right.$$

- Distance matrix

$$\text{objects}\left\{\overbrace{\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}}^{\text{objects}}\right.$$

Distance functions for real-valued vectors

- L_p norms or Minkowski distance:

$$L_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- **p = 1, L_1,** Manhattan (or city block) or Hamming distance:

$$L_1(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i| \right)$$

Distance functions for real-valued vectors

- L_p norms or Minkowski distance:

$$L_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- $p = 2$, L_2, Euclidean distance:

$$L_2(x, y) = \left(\sum_{i=1}^{d} (x_i - y_i)^2 \right)^{1/2}$$

Distance functions for real-valued vectors

- Dot product or cosine similarity

$$\cos(x, y) = \frac{x \cdot y}{||x||||y||}$$

- Can we construct a distance function out of this?
- When use the one and when the other?

Hamming distance for 0-1 vectors

x	0	1	0	0	1	0	0	1	0
y	1	0	0	0	0	1	0	1	1

$$L_1(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i| \right)$$

Hamming distance for 0-1 vectors

x	0	1	0	0	1	0	0	1	0
y	1	0	0	0	0	1	0	1	1

$$L_1(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i| \right)$$

How good is Hamming distance for 0-1 vectors?

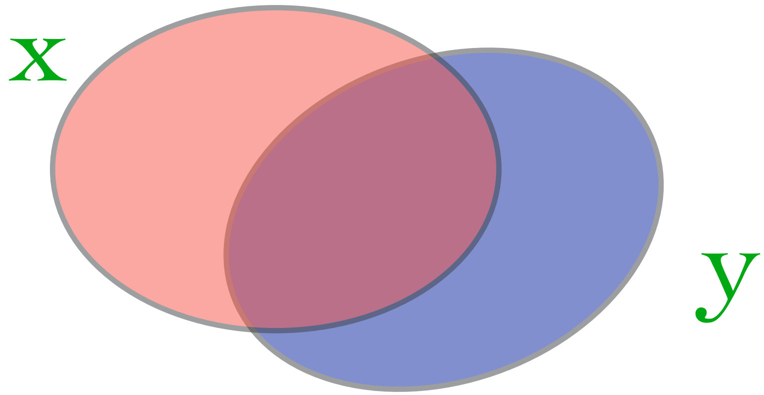
- **Drawback**

- Documents represented as sets (of words)
- Two cases
  - Two **very large** documents -- almost identical -- but for 5 terms
  - Two **very small** documents, with 5 terms each, disjoint

Distance functions for binary vectors or **sets**

- **Jaccard** similarity between binary vectors x and y (Range?)

$$\mathrm{JSim}(x,y) = \frac{|x \cap y|}{|x \cup y|}$$

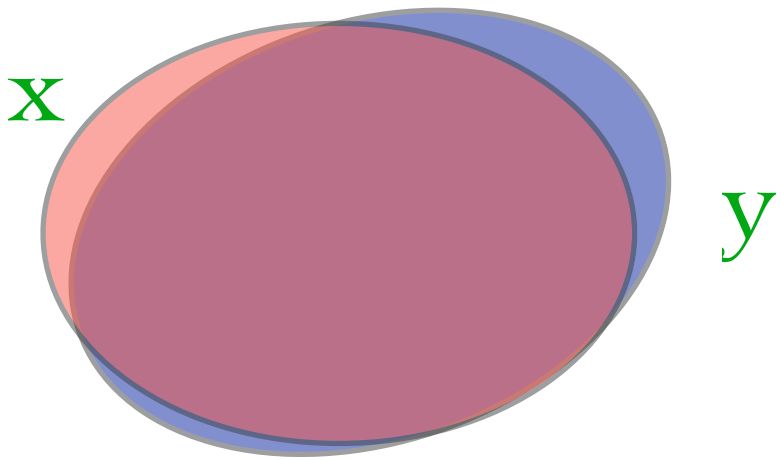


- **Jaccard** distance (Range?):

$$\mathrm{JDist}(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

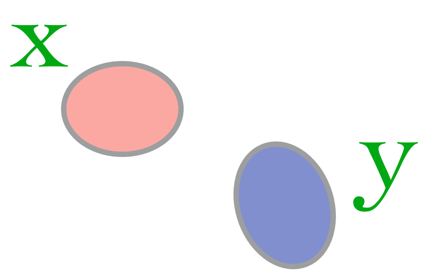
The previous example

- Case 1 (very large almost identical documents)



$J(x, y)$ almost 1

- Case 2 (small disjoint documents)



$J(x, y) = 0$

Jaccard similarity/distance

- Example:
  - JSim = 1/6
  - Jdist = 5/6

	Q1	Q2	Q3	Q4	Q5	Q6
X	1	0	0	1	1	1
Y	0	1	1	0	1	0