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)≥0** (non-negativity)
  - **d(i, i)=0** (isolation)
  - **d(i, j)= d(j, i)** (symmetry)
  - **d(i, j) ≤ d(i, h)+d(h, j)** (triangular inequality) [**Why do we need it?**]
- 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

attributes/dimensions

tuples/objects

$$\begin{bmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{il} & \cdots & x_{id} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nl} & \cdots & x_{nd} \end{bmatrix}$$

- Distance matrix

objects

objects

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

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) 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 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?):

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

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

Distance functions for strings

- **Edit distance** between two strings x and y is the **min** number of operations required to transform one string to another

- Operations: replace, delete, insert, transpose etc.

Examples of edit distance

- **Hamming distance** between two strings x and y of equal length is the number of positions in which the two strings differ from each other
- Examples: the Hamming distance between
  - "**toned**" and "**roses**" is 3.
  - 1011101 and 1001001 is 2.
  - 2173896 and 2233796 is 3.

Examples of edit distance

- **Edit distance** between two strings x and y of length n and m resp. is the **min** number of single-character edits (insertion, deletion, substitution) required to change one word to the other

Example

- I N T E N T I O N
- E X E C U T I O N

- I N T E * N T I O N
- * E X E C U T I O N
- d s s i s

Computing edit distance

- **Edit distance** is computed using **dynamic programming**

$$D(i,j) = \min\{D(i-1,j) + \mathrm{del}(X[i]),$$
$$D(i,j-1) + \mathrm{ins}(Y[j]),$$
$$D(i-1,j-1) + \mathrm{sub}(X[i],Y[j])\}$$

- Running time? Metric?