Distance measures for clustering

- What is the distance between two objects (rows in the data matrix)?
- (Earlier we talked about distances between sets of objects, given the distance between objects; now we look at the distance between objects.)
- Examples: objects are documents; what is the distance between two documents
- Objects are gene sequences (strings in ACGT); what is the distance?
- Euclidean distances vs. non-Euclidean distances (points in space vs. some other properties)

Desired properties of a distance measure

- Objects X, distance measure $d : X \times X \to R$
- Should be a metric
- $d(x, y) \geq 0$ for all x and y
- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z (triangle inequality)

Examples of metrics

- Assume points are in d-dimensional space
- $x = (x_1, x_2, \ldots, x_d)$ and $y = (y_1, y_2, \ldots, y_d)$
- Normal Euclidean distance (L_2 distance)

$$d_2(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}.$$

- This is the L_2 norm of $x - y$, also written as $||x - y||_2$
- L_p distance $p > 0$

$$d_p(x, y) = (\sum_{i=1}^{d} (x_i - y_i)^p)^(1/p)$$

- What is L_1 distance? What is L_∞ distance?

Distance and similarity

- Similarity $s(x, y) = M - d(x, y)$, where M is the maximal value the distance can obtain
- About equivalent concepts
- Not always

Other examples for vectors of 0s and 1s

- Jaccard distance: view $x = (x_1, x_2, \dots, x_d) \in \{0,1\}^d$ and $y = (y_1, y_2, \dots, y_d) \in \{0,1\}^d$ as sets
- $$J(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}.$$
- Cosine distance: view x and y as vectors, similarity between the vectors is measured as the cosine of the angle between the vectors
$$s(x, y) = \frac{x \cdot y}{||x||_2 ||y||_2}$$
How does this behave?

Distance notions for strings

- How close are two strings to each other?
- Edit distance
- Select a set of operations, say insert a character, delete a character, substitute one character for another
- $d(s, t)$: the smallest number of operations that is needed to transform string s into string t
- This is a metric (why?)
- Can be computed efficiently

Similarity for columns

- Similarity for columns of a 0-1 matrix M
- Jaccard coefficient for columns A and B:

$$J(A,B) = \frac{|\{t \in M | t(A) = 1 \wedge t(B) = 1\}|}{|\{t \in M | t(A) = 1 \vee t(B) = 1\}|}$$

- How many rows t of M have both A and B, divided by how many rows have at least one of A and B
- $|\{t \in M | t(A) = 1 \vee t(B) = 1\}| = f(A) + f(B) - f(AB)$
- Thus $J(A,B) = f(AB)/(f(A) + f(B) - f(AB))$
- High if A and B are strongly associated
- Alternatives: correlation between variables A and B

Other measures of similarity

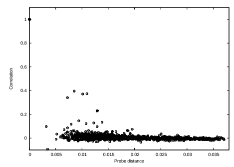
- Two columns A and B might actually be similar even if there are only a few rows with $A = B = 1$
- Example: retail data, two soft drinks A and B
- A and B are similar in the sense that the behavior of the customers who buy A is about the same as the behavior of the customers who buy B
- Can be very few customers who buy both A and B
- How to formalize this?

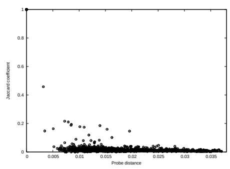
Probe similarity

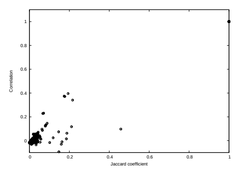
- Define the distance of A and B by the difference in the probabilities of the other variables
- $Pr(C|A=1)$: what fraction of the customers who buy A also buy C (conditional probability)
- $Pr(C|A=1) = f(AC)/f(A)$
- Let U be the set of attributes, and $U_{AB} = U \setminus \{A, B\}$
- $d(A,B) = \sum_{C \in U_{AB}} |Pr(C|A=1) - Pr(C|B=1)|$

G. Das, H. Mannila, P. Ronkainen, KDD 1998

Example: retail.over1000cols.01







Correlations between measures:

Probe	Corr	Jaccard
1.00000	-0.36364	-0.34927
-0.36364	1.00000	0.98819
-0.34927	0.98819	1.00000

Contextual similarity

- "Two words are similar, if they occur in similar sentences."
- "Two sentences are similar, if they contain similar words."
- How to implement this?