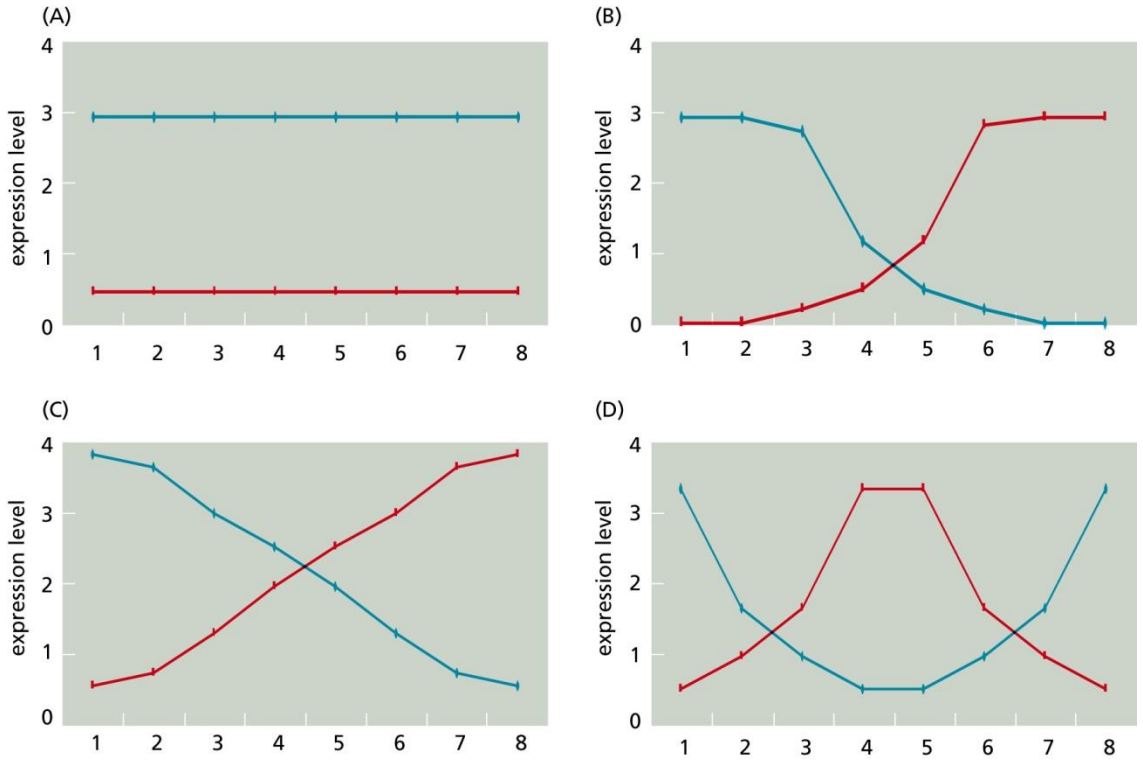
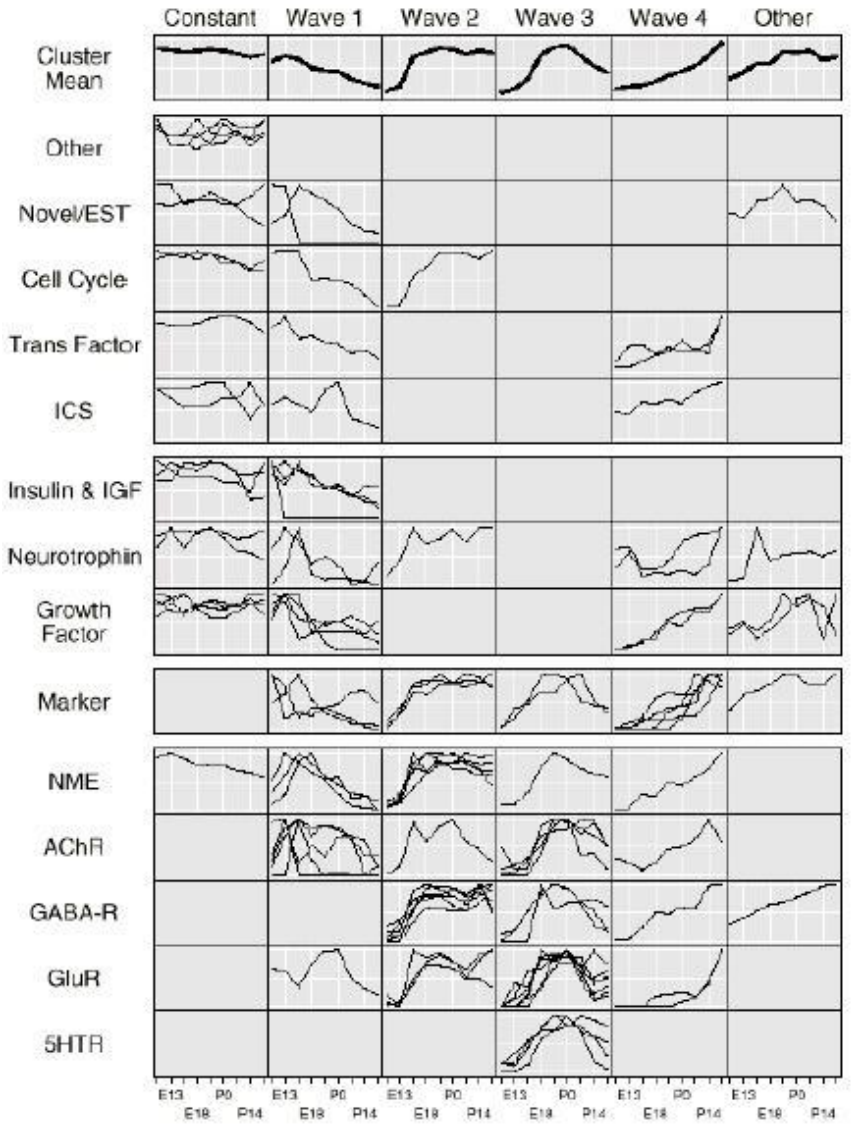
CISC 636 Computational Biology & Bioinformatics (Fall 2016)

Systems biology: Gene expressions profiling and clustering

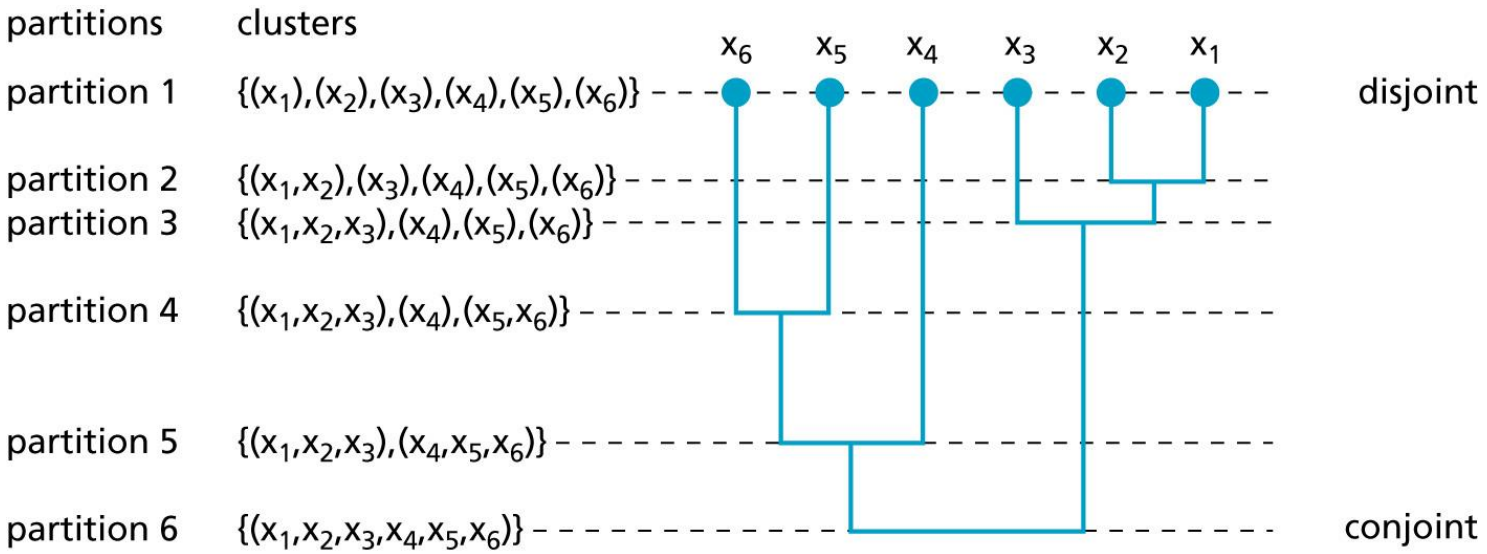
Typical expression profiles



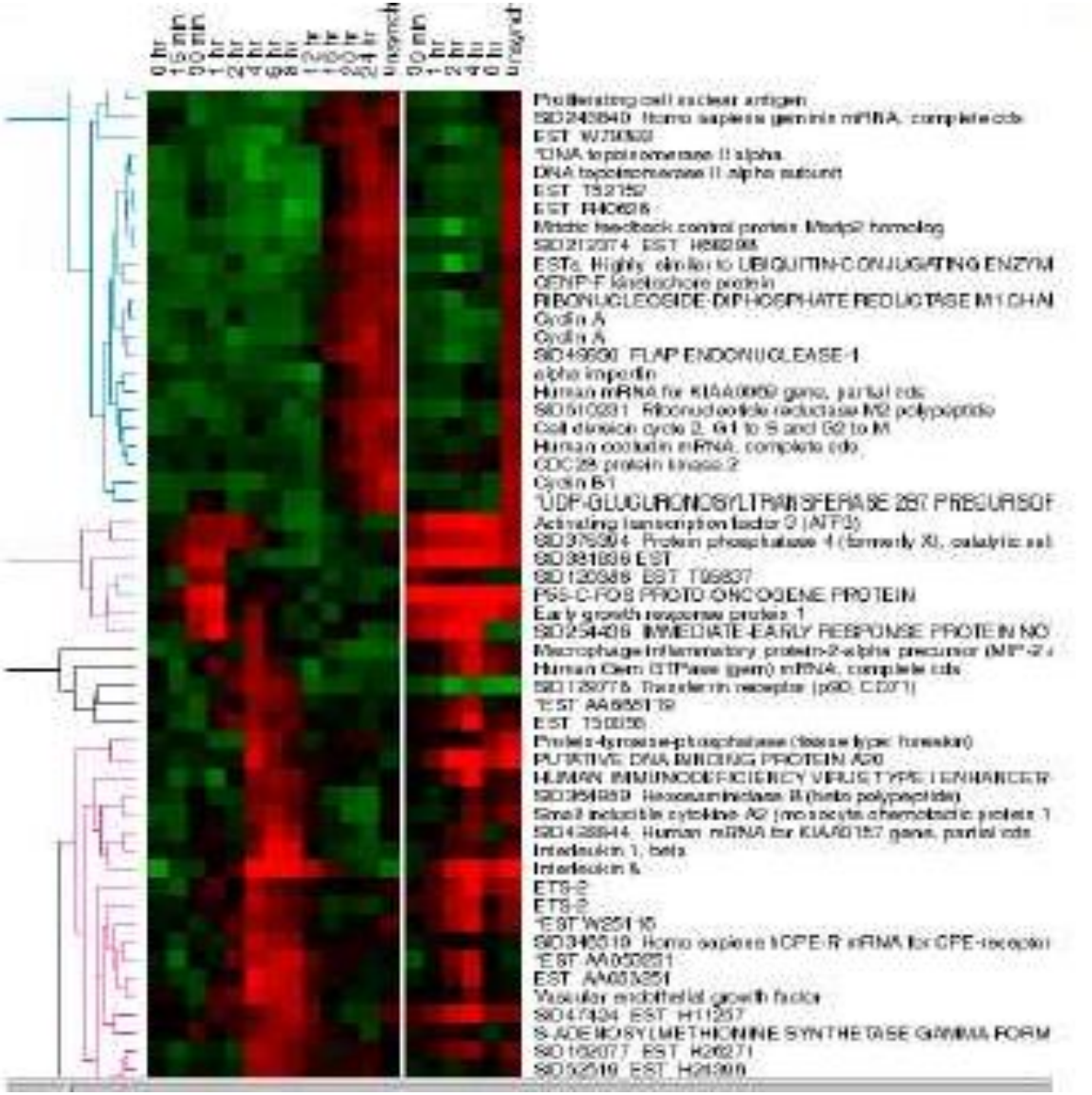

Constant
Wave 1
Wave 2
Wave 3
Wave 4
Other
Cluster Mean
Other
Novel/EST
Cell Cycle
Trans Factor
ICS
Insulin & IGF
Neurotrophin
Growth Factor
Marker
NME
AChR
GABA-R
GluR
5HTR
E13
P0
E18
P14

Russ Altman

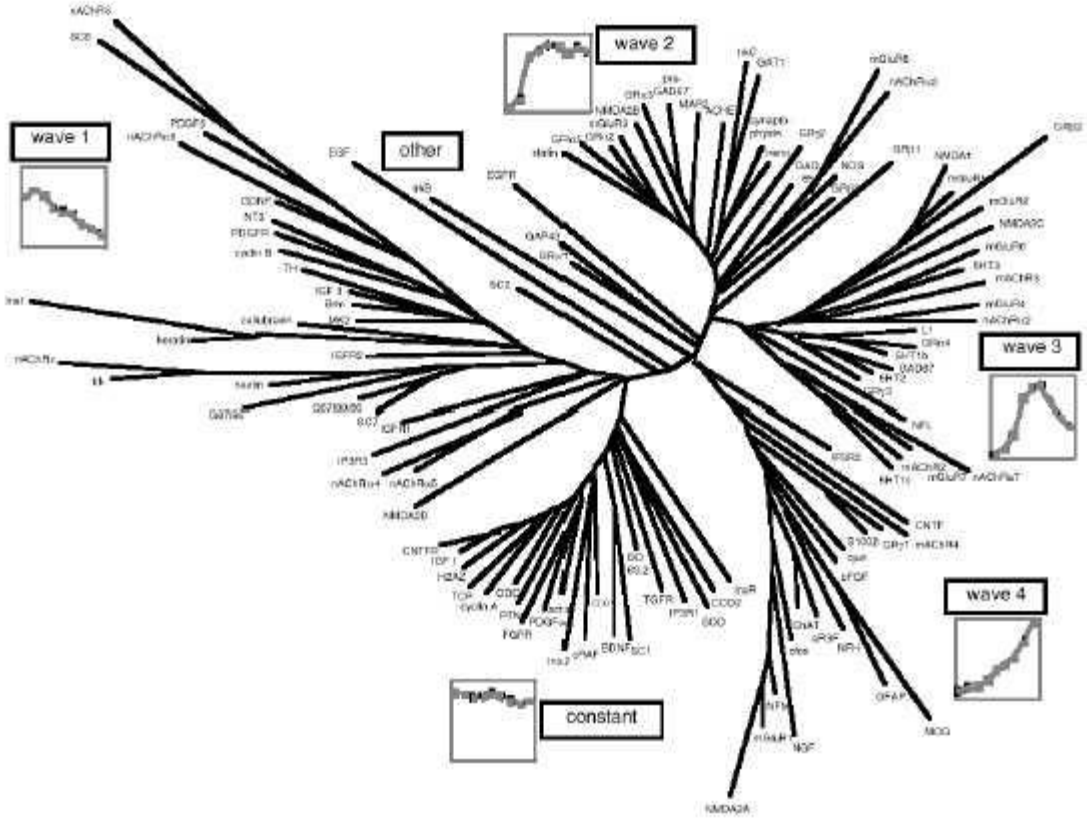
Hierarchical clustering



Hierarchical clustering

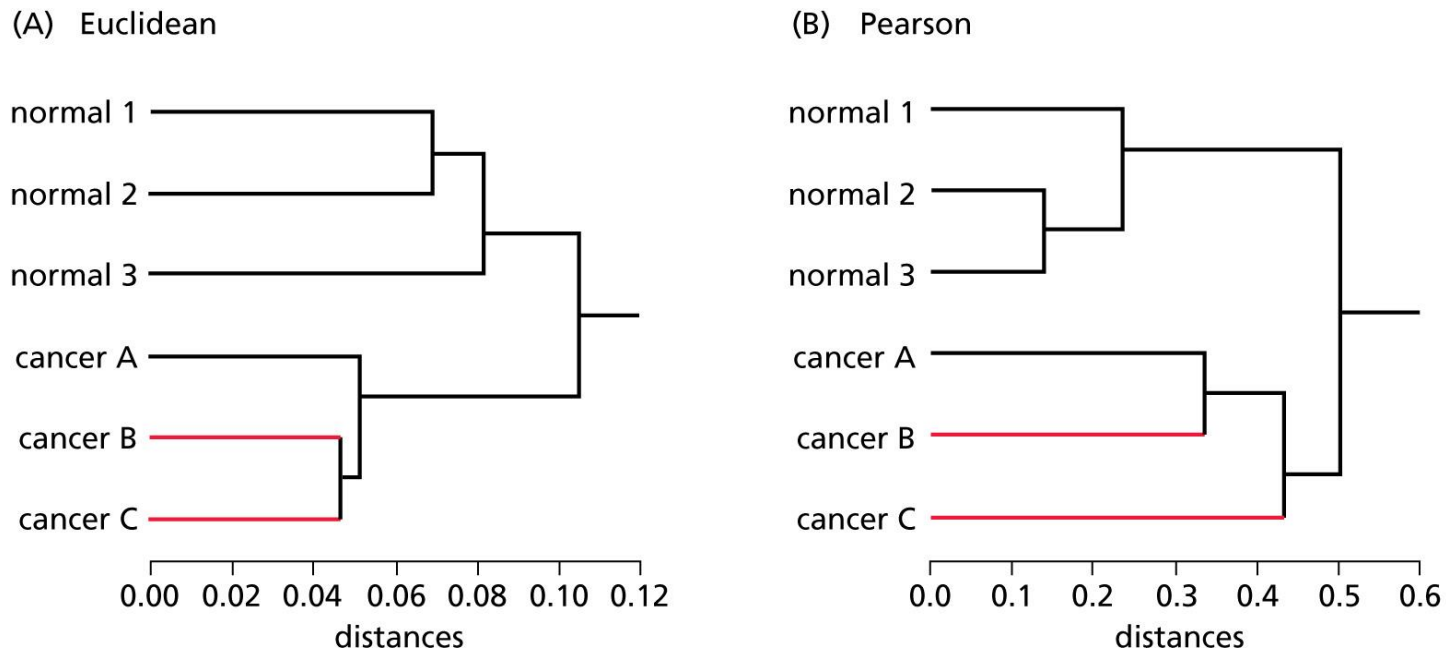


Russ Altman



Russ Altman

Effects of various metrics for measuring distance



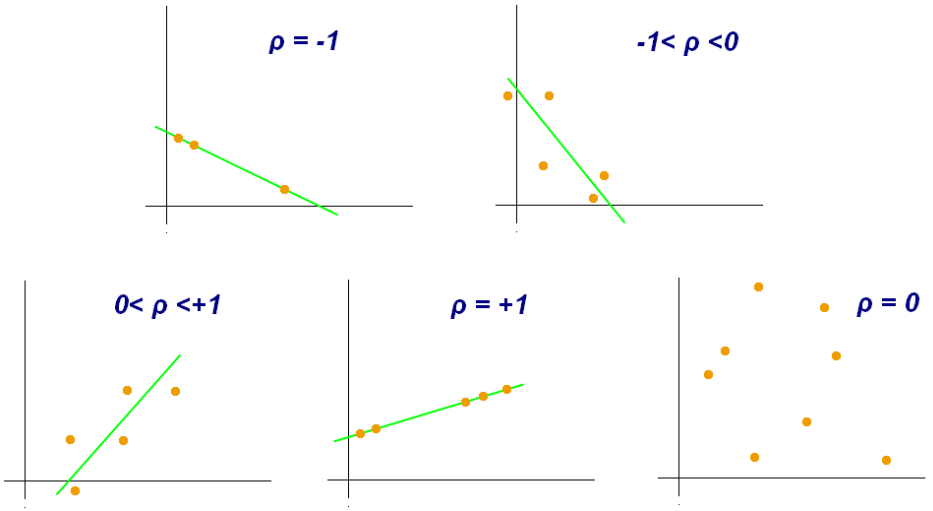
$$d_{x,y} = \sqrt{\sum (x_i - y_i)^2}$$

$$d_{X,Y} = 1 - \rho_{X,Y}.$$

Pearson correlation coefficient

For a population

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$



For a sample

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Pearson distance: $d_{X,Y} = 1 - \rho_{X,Y}.$

Effect of different clustering schemes

level

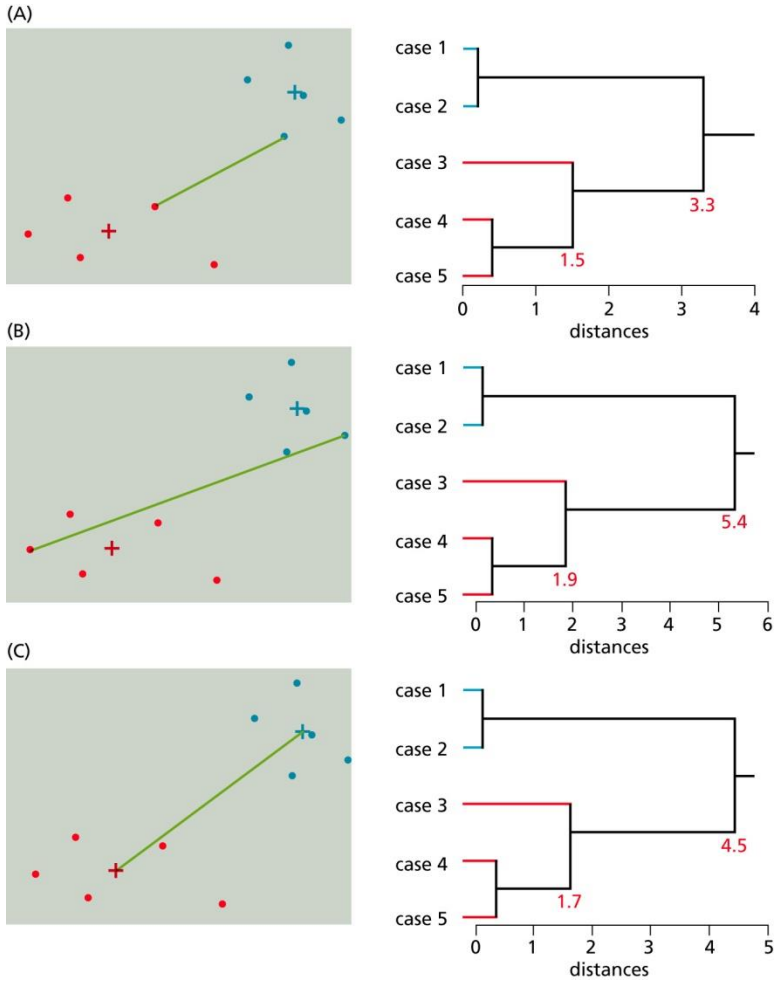
case1: 2.0

case 2: 2.2

case 3: 5.5

case 4: 7.0

case 5: 7.4



Single linkage

Complete linkage

centroid

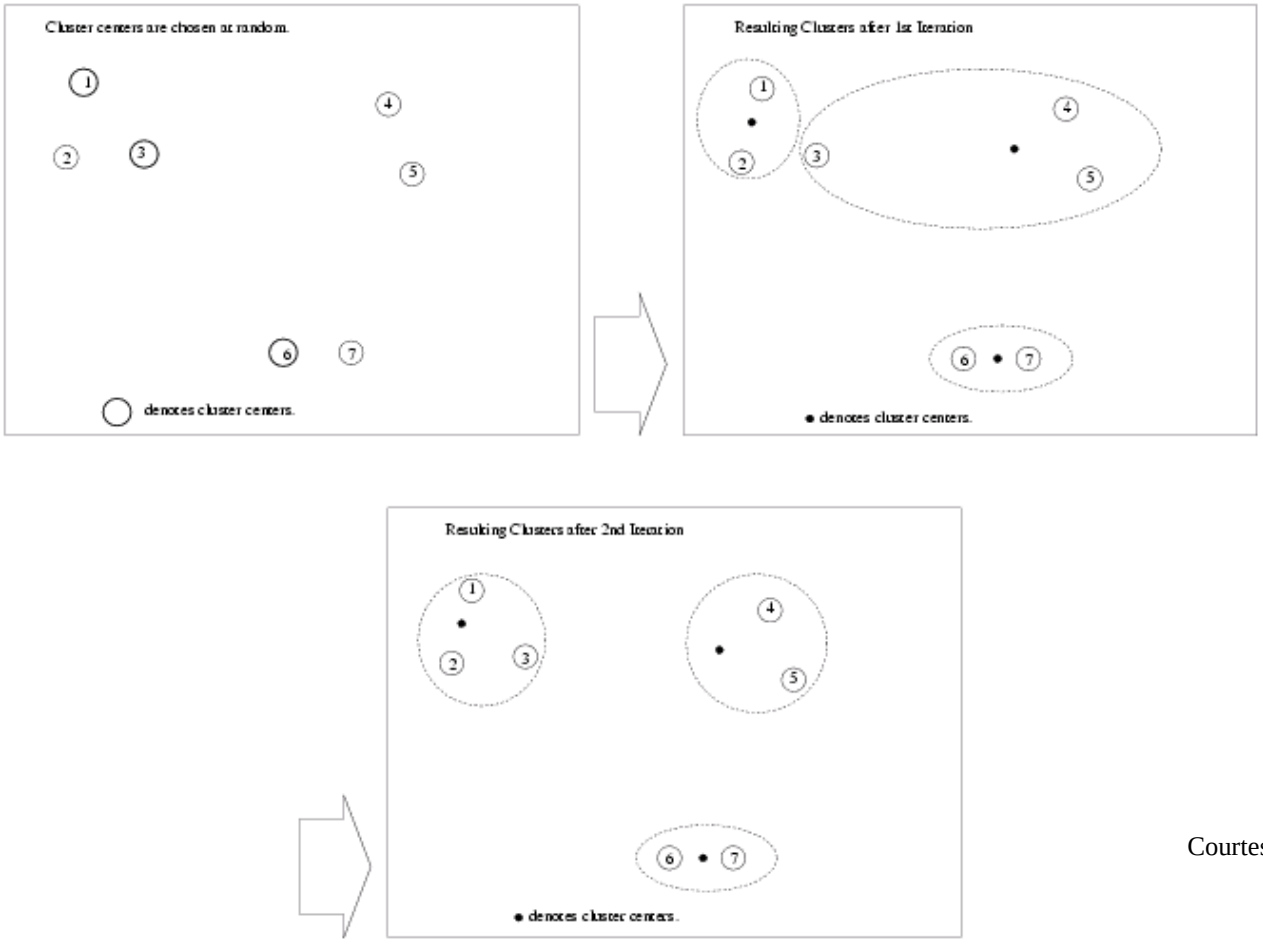
Iterative Distance-based Clustering (K-means)

Basic idea: Given a predetermined constant k (the number of clusters), iteratively recompute centers (means) of k clusters starting from randomly chosen k instances as centers.

1. K instances are chosen at random as cluster centers.
2. Instances are assigned to their closest cluster center, generating k cluster.
3. **while** (there is change in cluster centers)
4. Compute the centroid (mean) of all instances in each cluster.
5. Instances are assigned to their closest cluster center, generating k cluster.
6. **end**

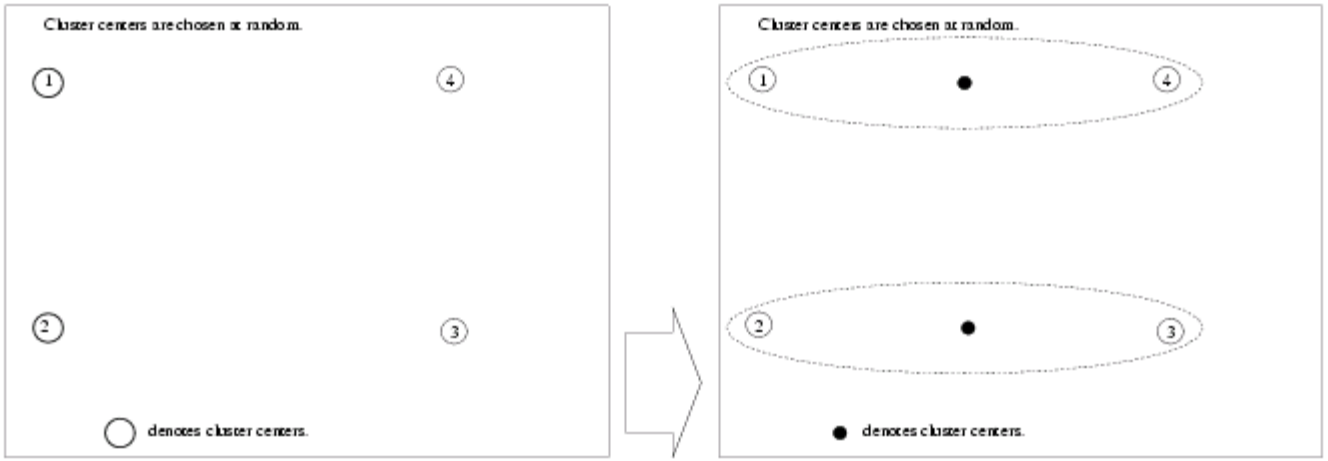
Courtesy of Sun Kim

A Correct Clustering Example



Courtesy of Sun Kim

An Incorrect Clustering Example



The initial choice of cluster centers, node 1 and node2, leads to an incorrect clustering. Obviously. a different choice of cluster centers, node 1 and node 3, result in a correct clustering.

Courtesy of Sun Kim

Discussion

1. The iterative procedure for k-means may end up with a local minimum, depending on the initial choice for cluster centers.
2. A simple heuristic is to run the k-mean clustering several times with different starting points.
3. How do we know the number of clusters in advance?
   Many different k can be tried.
4. K-mean clustering, as most clustering techniques, assumes that instances can be placed in Euclidian space.
5. Speeding up the K-mean algorithm is important.
   See the paper in SIGKDD Exploration (July 2000) by Farnstorm, Lewis, and Elkan.
   `http://www-cse.ucsd.edu/~elkan`

Courtesy of Sun Kim

Fuzzy k-means clustering

Fuzzy membership: Each data point **x** has some probability to belong to a cluster w (centered at **u**).

$$P(w|\mathbf{x})$$

The probabilities of cluster membership for each point are normalized

$$\sum_{i = 1 \text{ to } k} P(w_i|\mathbf{x}_j) = 1 \text{ for } j = 1, \ldots, n \qquad (1)$$

Cluster cost:

$$J = \sum_{i = 1 \text{ to } k} \sum_{j = 1 \text{ to } n} [P(w_i|\mathbf{x}_j)]^b \, ||\mathbf{x}_j - \mathbf{u}_i||^2. \qquad (2)$$

Condition for minimum cost:

$$\partial J / \partial \mathbf{u}_i = 0$$

$$\mathbf{u}_i = \left(\sum_{j=1 \text{ to } n} [P(w_i|\mathbf{x}_j)]^b \, \mathbf{x}_j\right) / \left(\sum_{j=1 \text{ to } n} [P(w_i|\mathbf{x}_j)]^b\right) \quad (3)$$

Update posterior probability as

$$P(w_i|\mathbf{x}_j) = (1/d_{ij})^{1/(b-1)} / \sum_{r=1 \text{ to } k} (1/d_{rj})^{1/(b-1)} \quad (4)$$

where $d_{ij} = \|\mathbf{x}_j - \mathbf{u}_i\|^2$.

Fuzzy k-means clustering algorithm

initialize $\mathbf{u}_1, \ldots, \mathbf{u}_k$

  normalize $P(w_i|\mathbf{x}_j)$ by eq(1)

  do recompute $\mathbf{u}_i$ for i = 1 to k by eq(3)

    recompute $P(w_i|\mathbf{x}_j)$ by eq(4)

  until small change in $\mathbf{u}_i$ and $P(w_i|\mathbf{x}_j)$

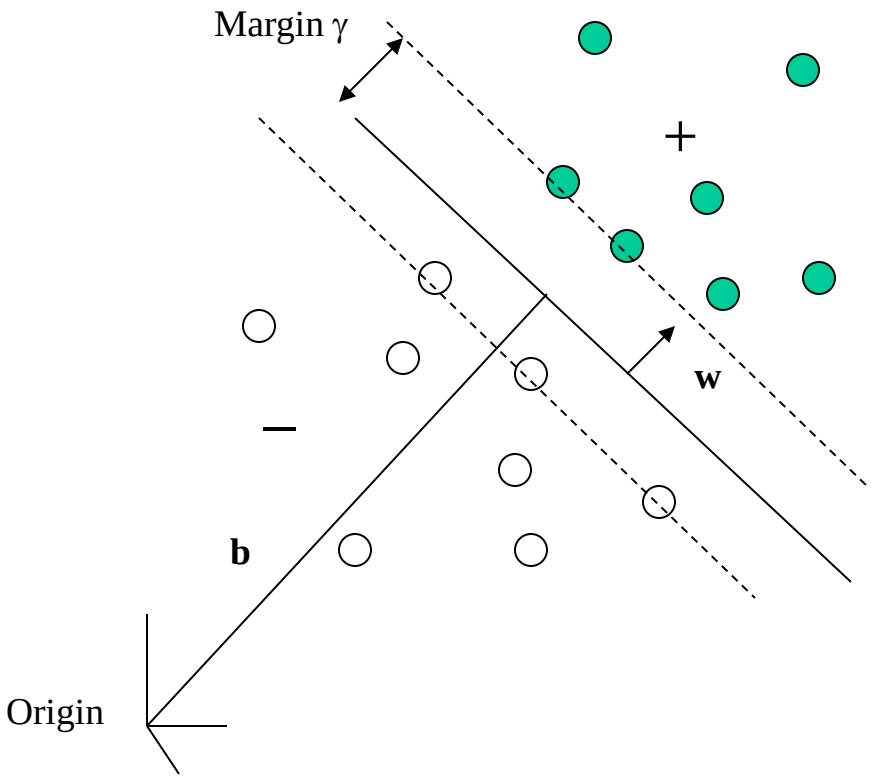
return $\mathbf{u}_1, \ldots, \mathbf{u}_k$.

Classical k-means is a special case when membership is defined as

$P(w_i|\mathbf{x}_j) = 1$ if $\|\mathbf{x}_j - \mathbf{u}_i\| < \|\mathbf{x}_j - \mathbf{u}_{i'}\|$ for all $i' \neq i$.

$= 0$ otherwise.

Support vector machine (SVM)



$\mathbf{w} \cdot \mathbf{x}_i + b > 0$ if $y_i = +1$

$\mathbf{w} \cdot \mathbf{x}_i + b < 0$ if $y_i = -1$

An unknown x is classified as sign($\mathbf{w} \cdot \mathbf{x} + b$)

Application of SVM classification