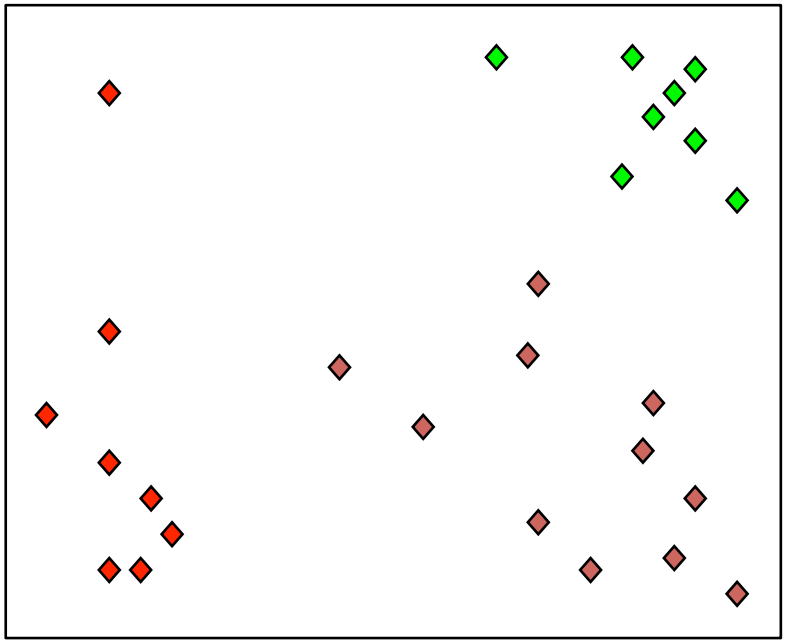
Introduction to Machine Learning: Clustering Algorithms

02-223 How to Analyze Your Own Genome

Fall 2013

What is Clustering?

- Organizing data into *clusters* such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects.



What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

Webster's Dictionary

Similarity is hard to define, but… "*We know it when we see it*"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Defining Distance Measures

Definition: Let x and *y* be two objects from the universe of possible objects. The distance (dissimilarity) between *x* and *y* is a real number denoted by *D*(*x*,*y*)

A few examples:

- Euclidian distance $$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

$x = (x_1, x_2, \ldots, x_J), y = (y_1, y_2, \ldots, y_J)$

J: the number of features

- Correlation coefficient

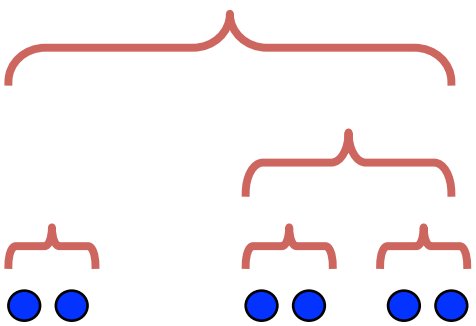
$$s(x,y) = \frac{\sum_i^J (x_i - \mu_x)(y_i - \mu_y)}{(J-1)\sigma_x \sigma_y}$$

Clustering Algorithms

- Hierarchical agglomerative clustering
- K-means clustering algorithm
- Gaussian mixture model

Hierarchical Clustering

- Probably the most popular clustering algorithm in computational biology

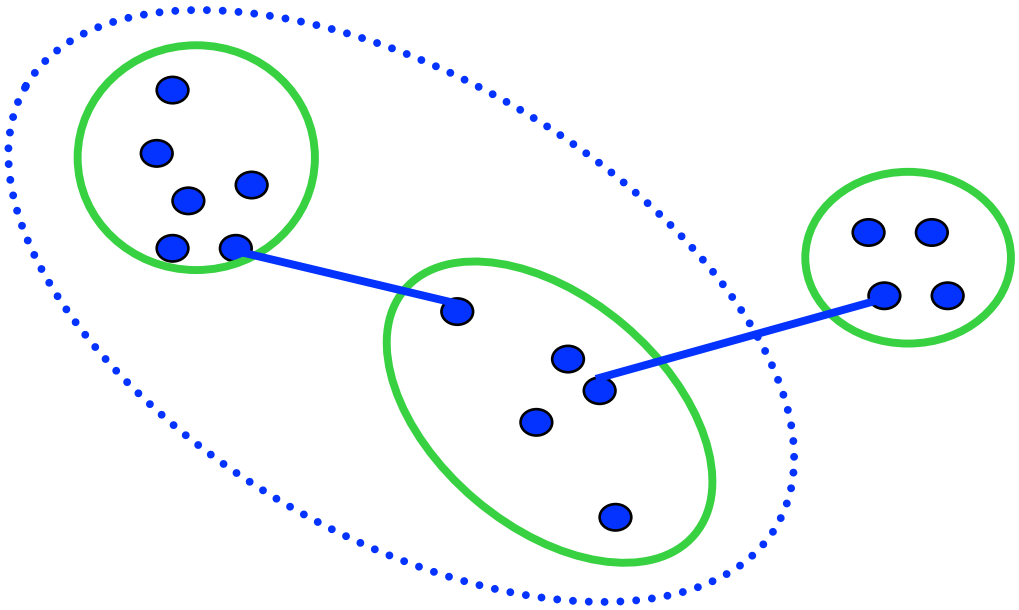


dendrogram

- Agglomerative (bottom-up)
- Algorithm:
  1. Initialize: each item a cluster
  2. Iterate:
     - select two most *similar* clusters
     - merge them
  3. Halt: when there is only one cluster left

Similarity Criterion: Single Linkage

- cluster similarity = similarity of two most similar members



- Potentially long and skinny clusters

In most cases (1-r^2), where r^2 is the correlation coefficient, is used as similarity measure between samples

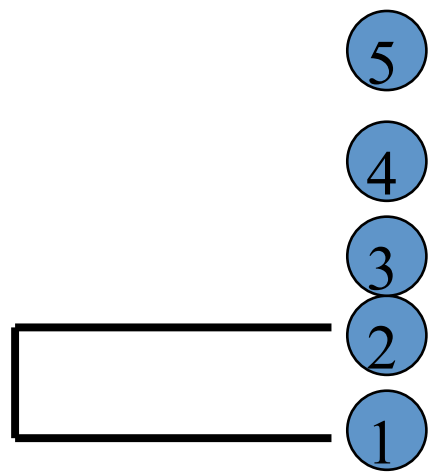
Example: Single Linkage

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2),3} = \min\{d_{1,3}, d_{2,3}\} = \min\{6,3\} = 3$$

$$d_{(1,2),4} = \min\{d_{1,4}, d_{2,4}\} = \min\{10,9\} = 9$$

$$d_{(1,2),5} = \min\{d_{1,5}, d_{2,5}\} = \min\{9,8\} = 8$$

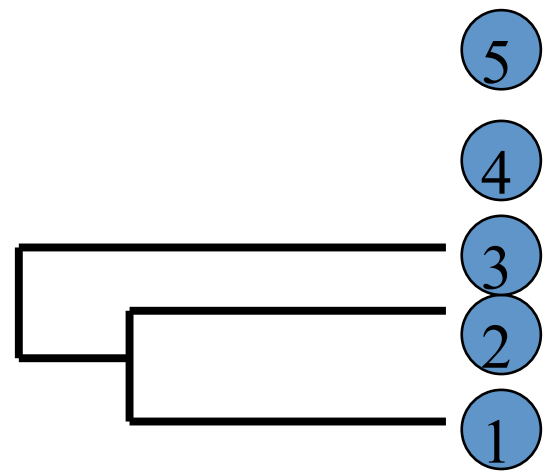


Example: Single Linkage

$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0 & & & & \\
2 & 2 & 0 & & & \\
3 & 6 & 3 & 0 & & \\
4 & 10 & 9 & 7 & 0 & \\
5 & 9 & 8 & 5 & 4 & 0
\end{array}
\quad \Rightarrow \quad
\begin{array}{c|cccc}
 & (1,2) & 3 & 4 & 5 \\
\hline
(1,2) & 0 & & & \\
3 & 3 & 0 & & \\
4 & 9 & 7 & 0 & \\
5 & 8 & 5 & 4 & 0
\end{array}
\quad \Rightarrow \quad
\begin{array}{c|ccc}
 & (1,2,3) & 4 & 5 \\
\hline
(1,2,3) & 0 & & \\
4 & 7 & 0 & \\
5 & 5 & 4 & 0
\end{array}
$$

$$d_{(1,2,3),4} = \min\{d_{(1,2),4}, d_{3,4}\} = \min\{9,7\} = 7$$

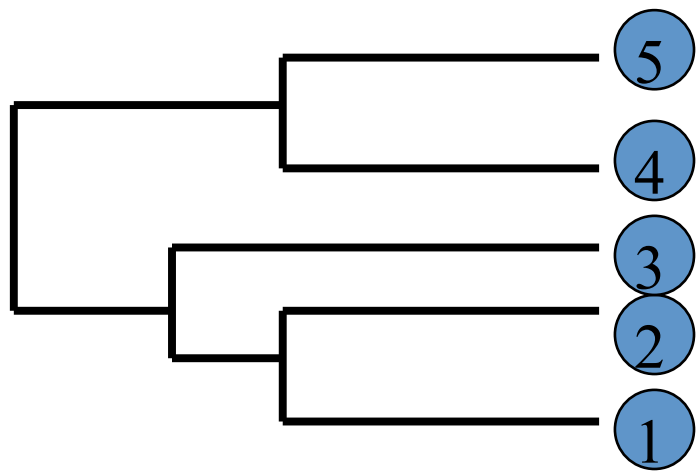
$$d_{(1,2,3),5} = \min\{d_{(1,2),5}, d_{3,5}\} = \min\{8,5\} = 5$$



Example: Single Linkage

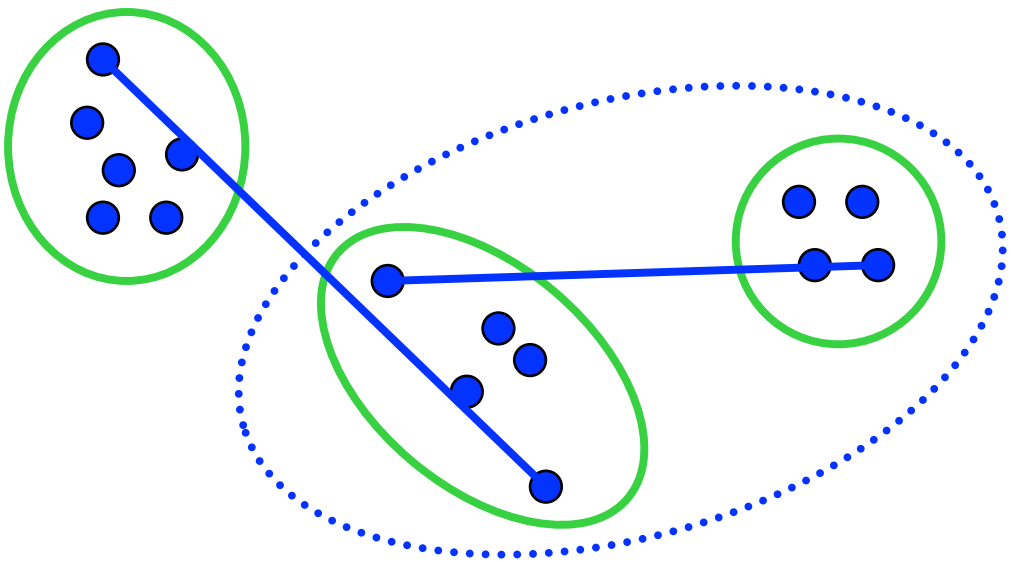
$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 2 & 0 & & & \\ 3 & 6 & 3 & 0 & & \\ 4 & 10 & 9 & 7 & 0 & \\ 5 & 9 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ \hline (1,2) & 0 & & & \\ 3 & 3 & 0 & & \\ 4 & 9 & 7 & 0 & \\ 5 & 8 & 5 & 4 & 0 \end{array} \Rightarrow \begin{array}{c|ccc} & (1,2,3) & 4 & 5 \\ \hline (1,2,3) & 0 & & \\ 4 & 7 & 0 & \\ 5 & 5 & 4 & 0 \end{array}$$

$$d_{(1,2,3),(4,5)} = \min\{d_{(1,2,3),4}, d_{(1,2,3),5}\} = 5$$



Similarity Criterion: Complete Linkage

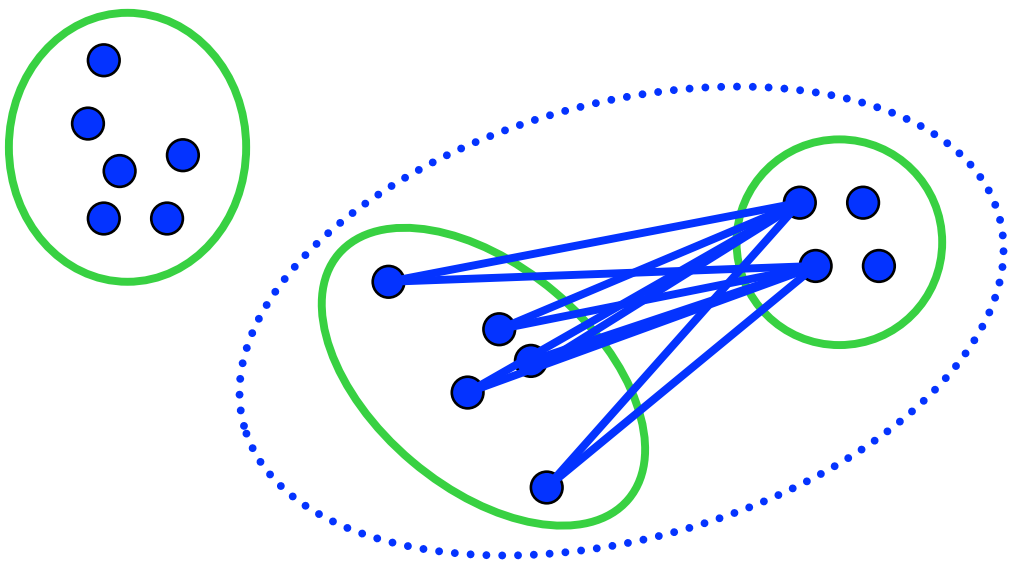
- cluster similarity = similarity of two least similar members



+ tight clusters

Similarity Criterion: Average Linkage

- cluster similarity = average similarity of all pairs

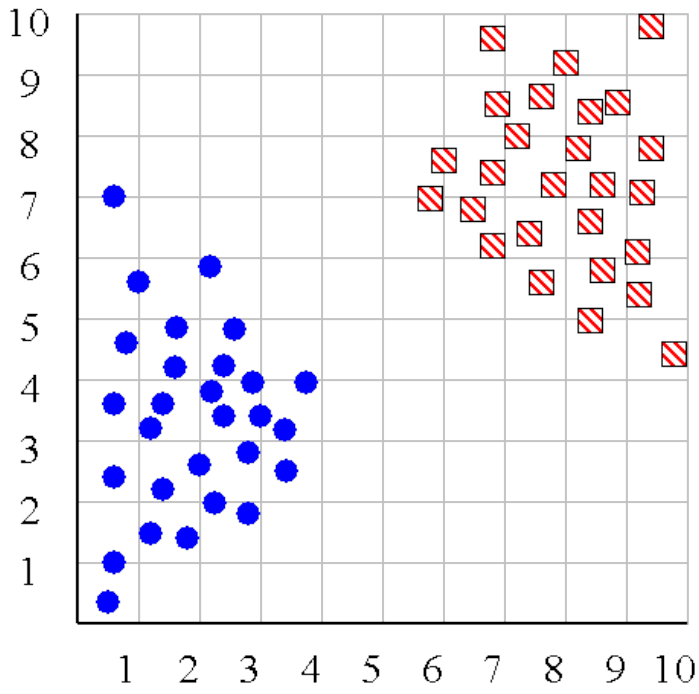


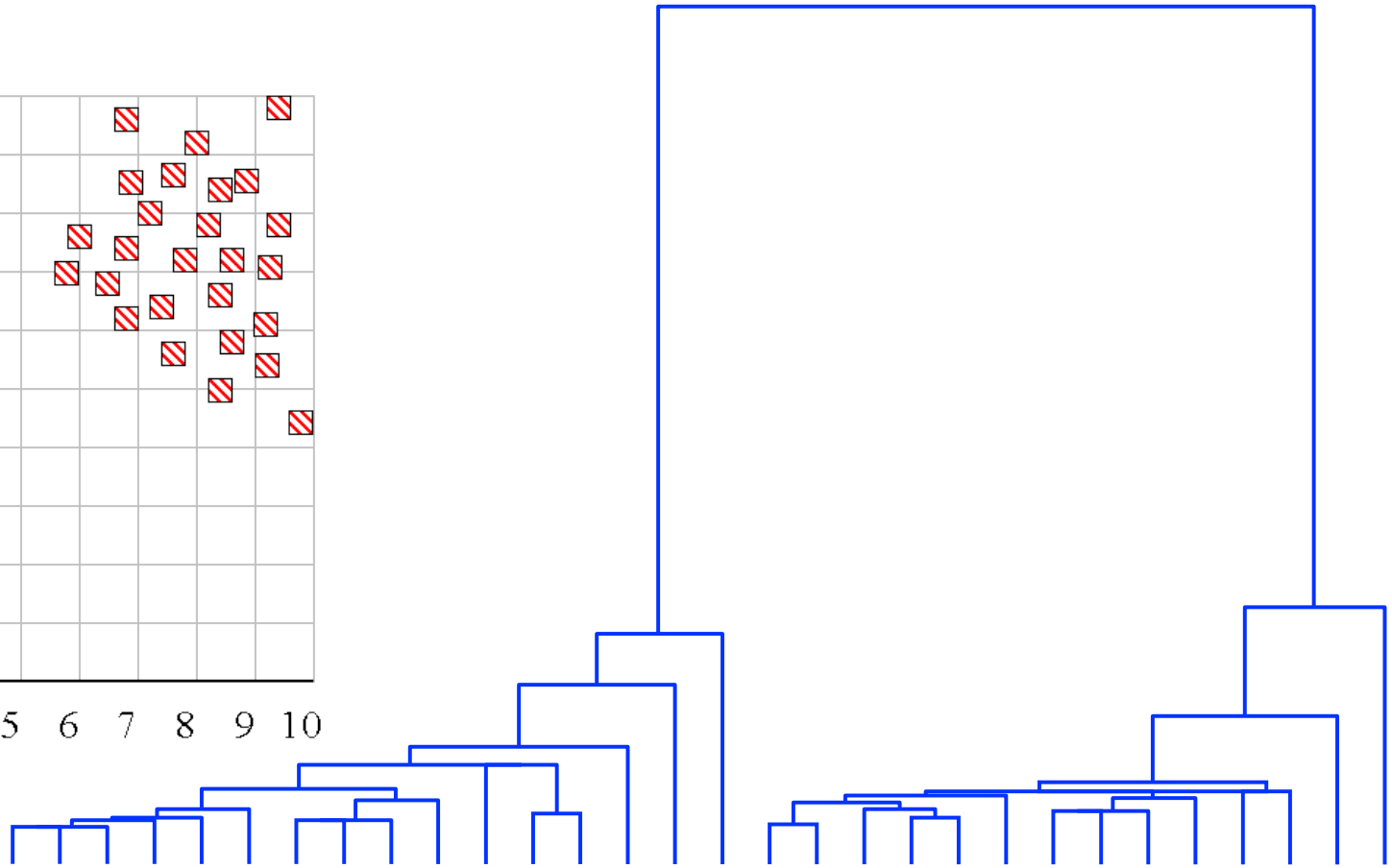
the most widely used similarity measure

Robust against noise

But What Are the Clusters?

In some cases we can determine the “correct” number of clusters. However, things are rarely this clear cut, unfortunately.





K-means Clustering Algorithm: Partitional Clustering

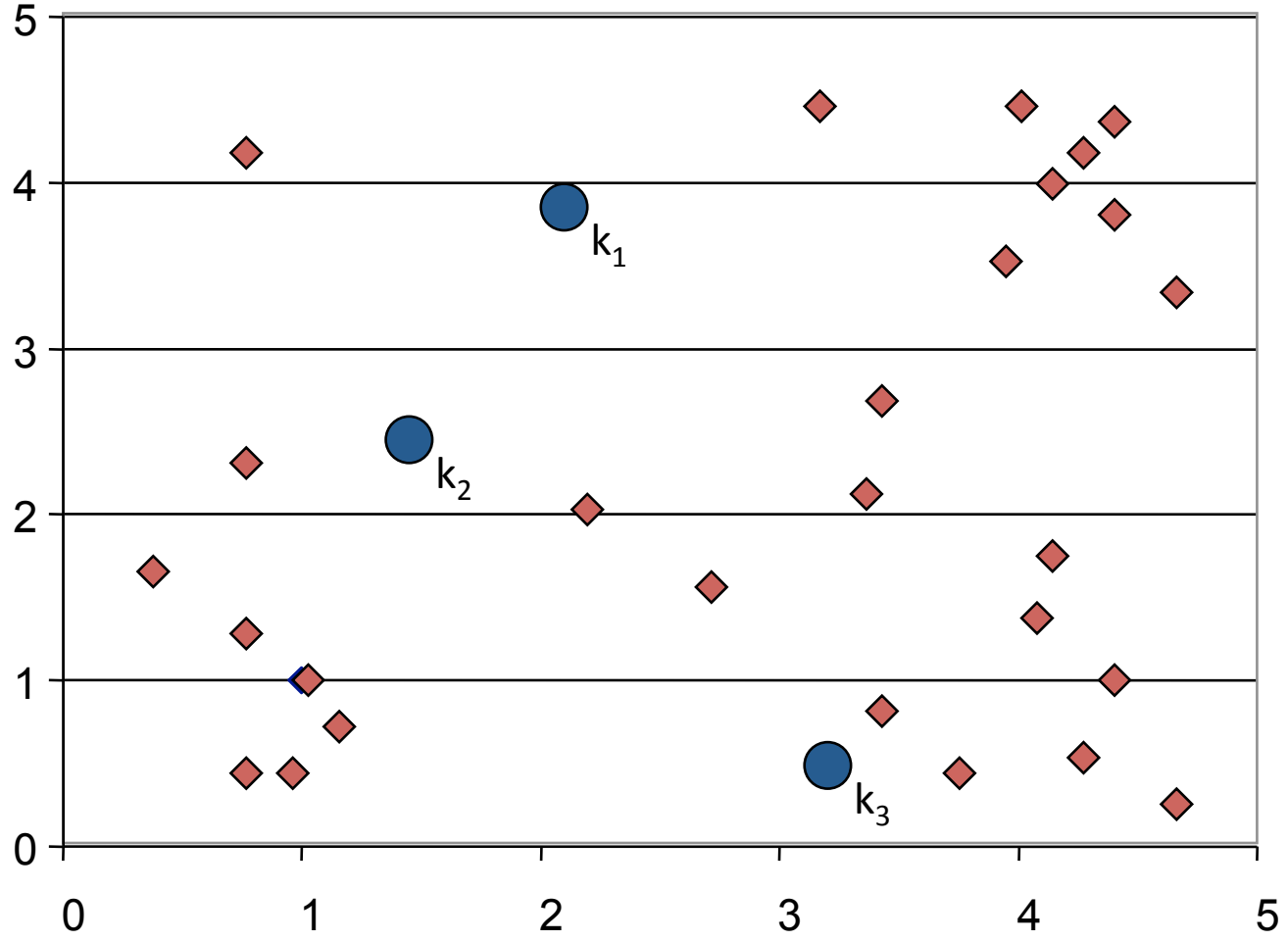
- Nonhierarchical, each object is placed in exactly one of K non-overlapping clusters.

- the user has to specify the desired number of clusters K.

- In hierarchical clustering, we use similarity measures between two observed samples, whereas in K-means clustering, we use the similarity measures between an observed sample and the cluster center (mean).

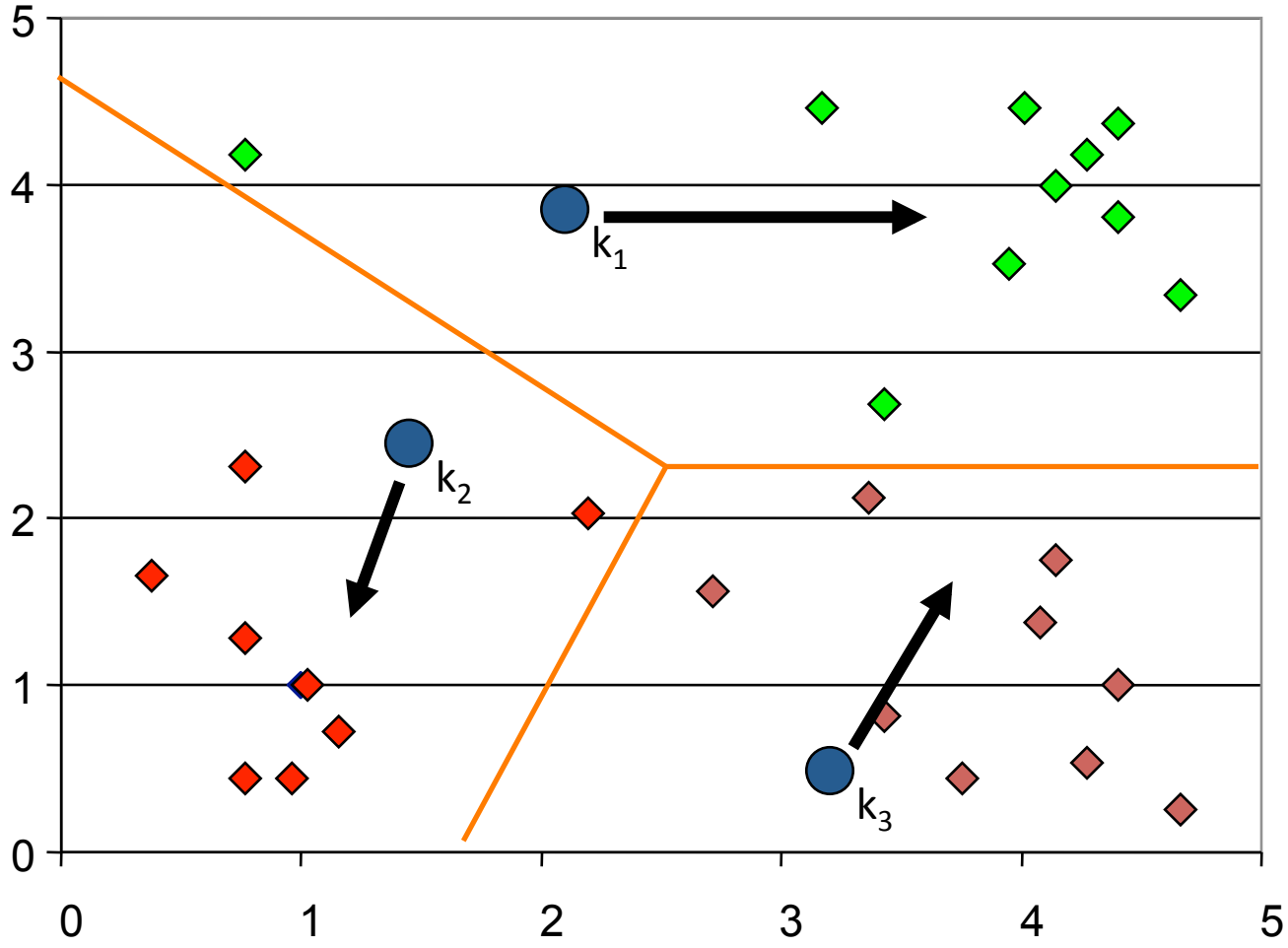
K-means Clustering: Initialization

- For a pre-defined number of clusters *K*, initialize *K* centers randomly



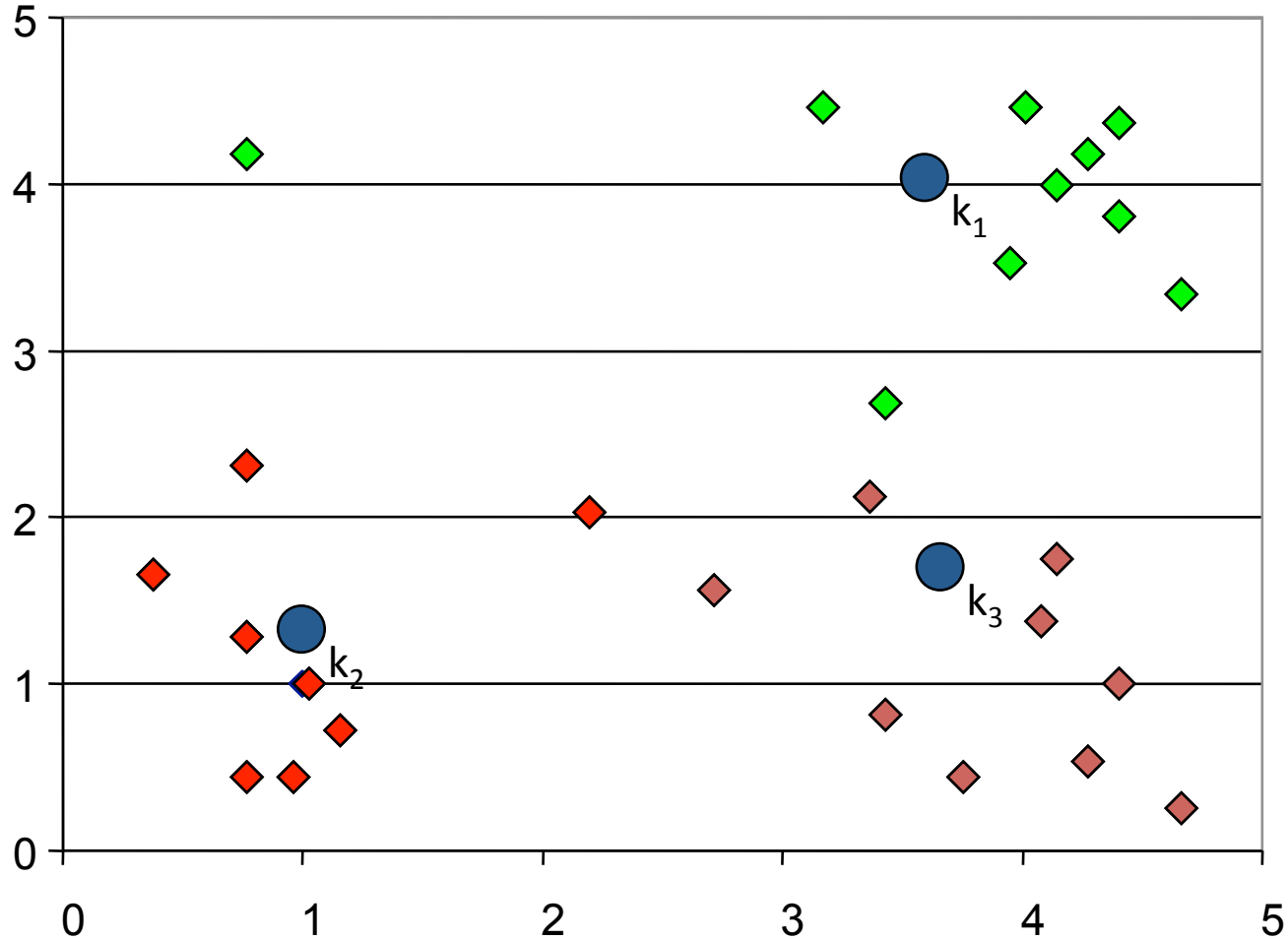
K-means Clustering: Iteration 1

- Iterate between the following two steps
  - Assign all objects to the nearest center.
  - Move a center to the mean of its members.



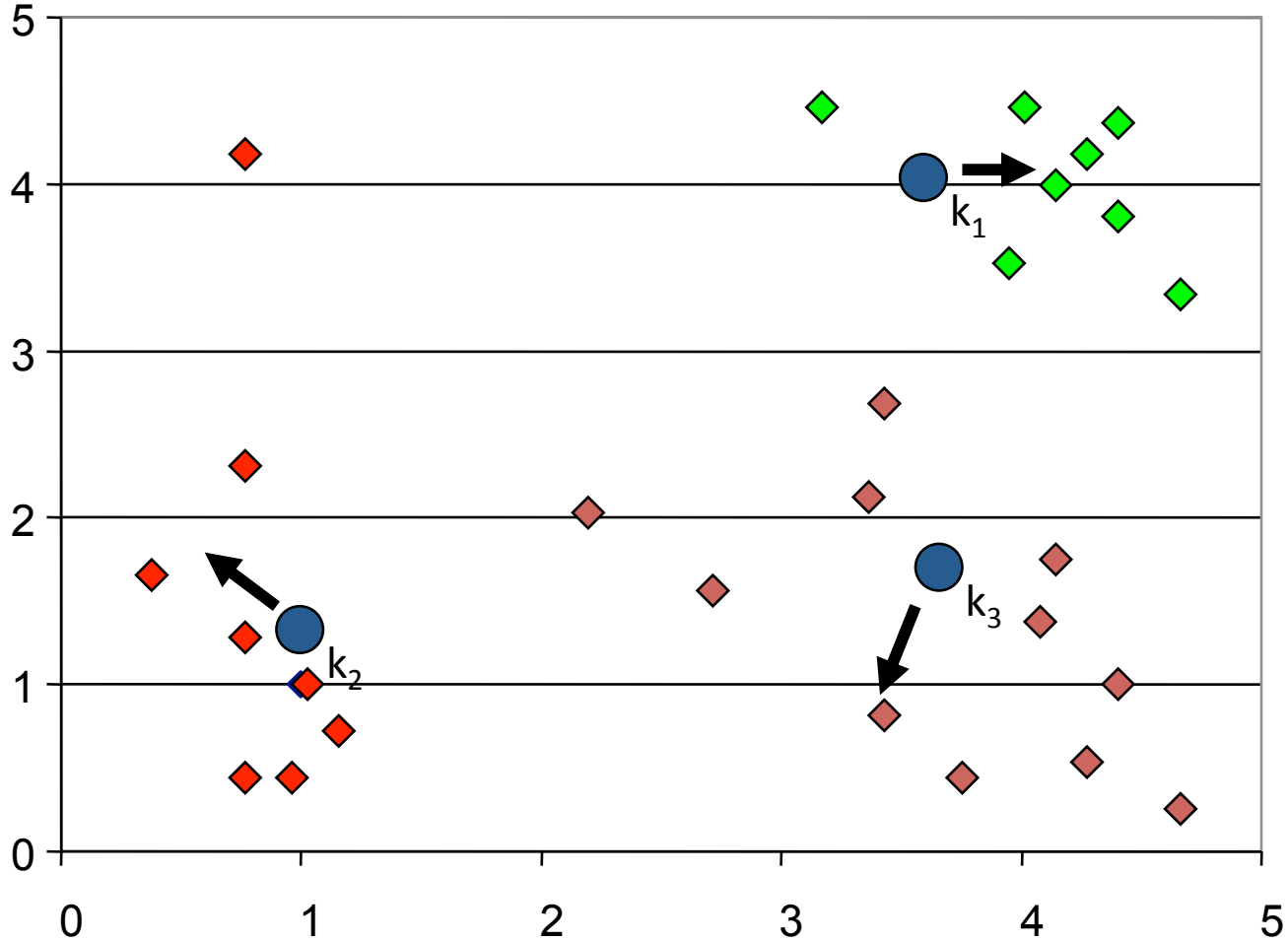
K-means Clustering: Iteration 2

- After moving centers, re-assign the objects…



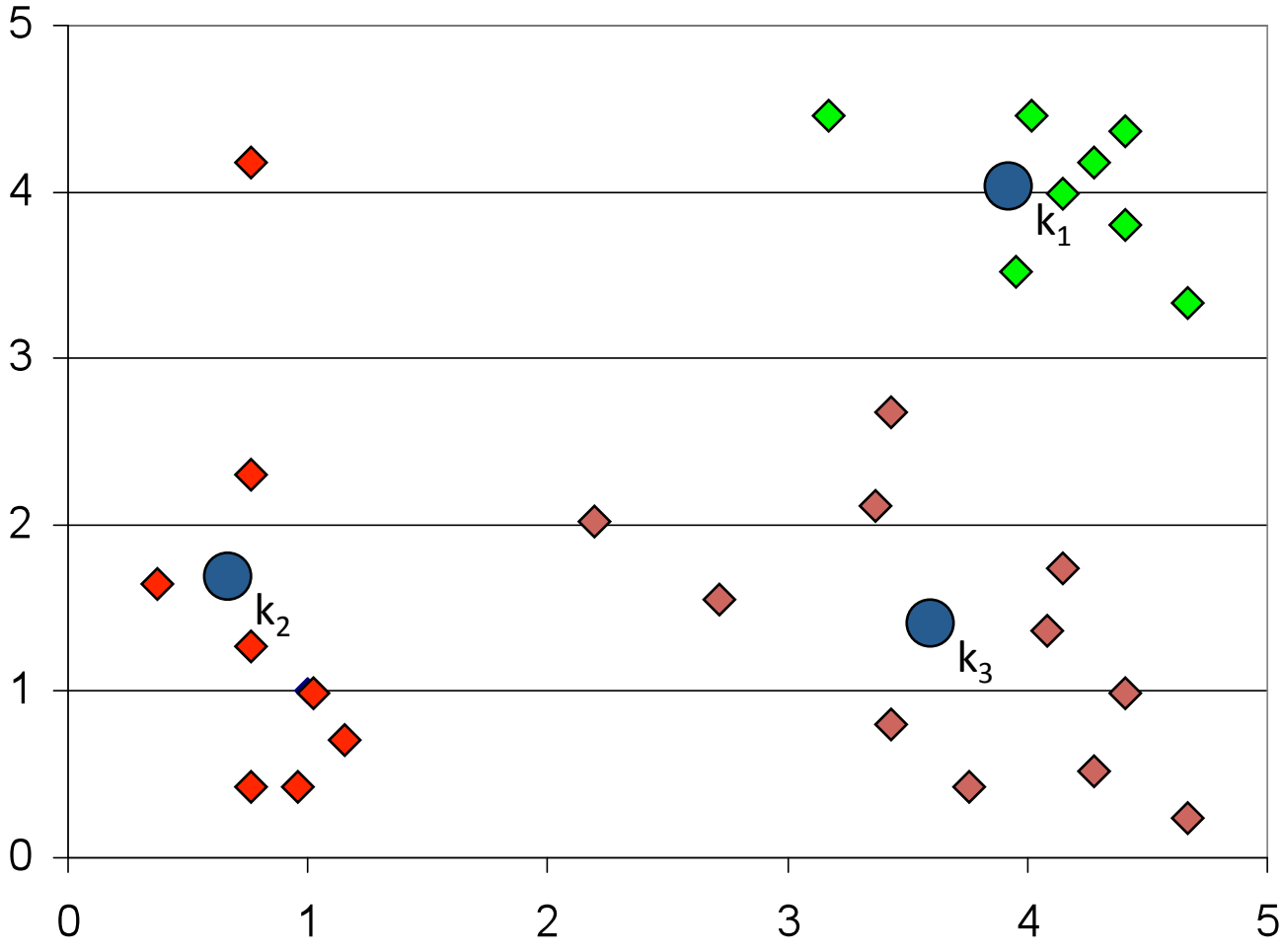
K-means Clustering: Iteration 2

- After moving centers, re-assign the objects to nearest centers.
- Move a center to the mean of its new members.



K-means Clustering: Finished!

- Re-assign and move centers, until no objects changed membership.



Algorithm *k-means*

1. Decide on a value for K, the number of clusters.

2. Initialize the K cluster centers randomly.

3. Decide the cluster memberships of the N objects by assigning them to the nearest cluster center.

4. Re-estimate the K cluster centers, by assuming the memberships found above are correct.

5. Repeat 3 and 4 until none of the N objects changed membership in the last iteration.

Algorithm *k-means*

1. Decide on a value for *K*, the number of clusters.

2. Initialize the *K* cluster centers (randomly, if necessary).

3. Decide the cluster memberships of the *N* objects by assigning them to the nearest cluster center.

4. Re-estimate the *K* cluster centers, by assuming the memberships found above are correct.

5. Repeat 3 and 4 until none of the *N* objects changed membership in the last iteration.

Use one of the distance / similarity functions we discussed earlier

Average / median of cluster members

Gaussian Mixture Model as Soft-Clustering

- In K-means algorithm, each sample can be assigned to only a single cluster.

- Gaussian mixture models relax this assumption
  - Each sample can be assigned to multiple clusters with certain probabilities.
  - The data for each cluster is modeled with a Gaussian distribution
    - Mean parameter η: cluster center
    - Variance parameter σ^2: cluster width

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\eta)^2}{2\sigma^2}}$$

Soft-Clustering of Individuals into Three Clusters with Gaussian Mixture Model

Probability of	Cluster 1	Cluster 2	Cluster 3	Sum
Individual 1	0.1	0.4	0.5	1
Individual 2	0.8	0.1	0.1	1
Individual 3	0.7	0.2	0.1	1
Individual 4	0.10	0.05	0.85	1
Individual 5	…	…	…	1
Individual 6	…	…	…	1
Individual 7	…	…	…	1
Individual 8	…	…	…	1
Individual 9	…	…	…	1
Individual 10	…	…	…	1

- Each individual can assigned to more than one clusters with a certain probability.
- For each individual, the probabilities for all clusters should sum to 1. (i.e., each row should sum to 1.)
- Each cluster is explained by a cluster center variable (i.e., cluster mean)

Summary

- Clustering algorithms
  - Hierarchical agglomerative clustering
    - Build a dendrogram over samples using similarity measures
  - K-means clustering algorithm
    - Partition samples into K clusters
    - Hard assignment: each sample can belong to only one cluster
  - Gaussian mixture models for clustering
    - Statistical/Probabilistic variation of K-means algorithm
    - Soft assignment: each sample can belong to multiple clusters and this uncertainty in cluster assignment is represented as probabilities that sum to 1.