CLUSTERING

CLUSTERING

- This is our first unsupervised learning technique

  Q: What is unsupervised learning again?

CLUSTERING

- This is our first unsupervised learning technique

  Q: What is unsupervised learning again?

  A: No outputs given. Trying to find structure

CLUSTERING

- This is our first unsupervised learning technique

Q: What is unsupervised learning again?

A: No outputs given. Trying to find structure

Q: What is clustering?

CLUSTERING

- This is our first unsupervised learning technique

Q: What is unsupervised learning again?

A: No outputs given. Trying to find structure

Q: What is clustering?

A: Grouping similar data points together. (clusters)

CLUSTERING

- This is our first unsupervised learning technique

Q: What is unsupervised learning again?

A: No outputs given. Trying to find structure

Q: What is clustering?

A: Grouping similar data points together. (clusters)

Want to find structure within the data

K-MEANS CLUSTERING

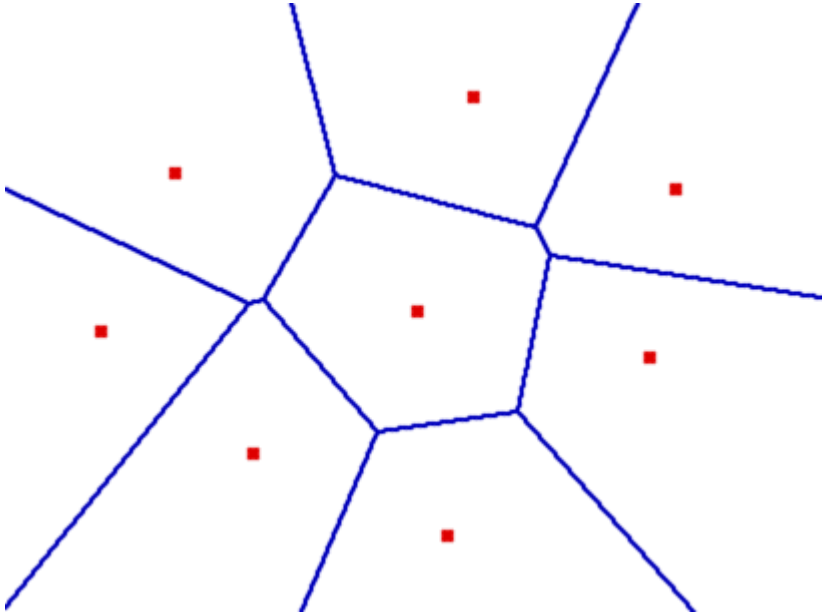
K-MEANS

Q: What is K-Means?

K-MEANS

Q: What is K-Means?

A: A greedy algorithm that partitions the data into k clusters



http://www.dma.fi.upm.es/mabellanas/tfcs/fvd/images/voronoi.gif

K-MEANS

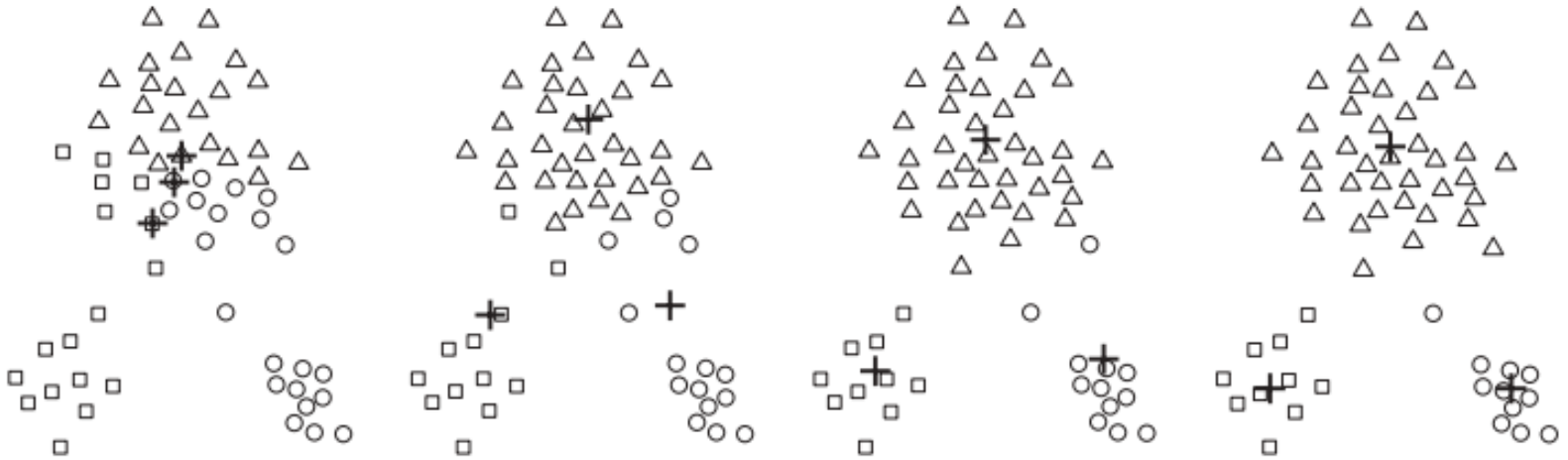
K-Means Algorithm

Input: k

1. Choose k initial starting positions. Mean of the cluster
2. Assign each point to its nearest mean
3. Recalculate mean = average of assigned points
4. Repeat steps 2-3 until convergence

Running time is k*n*d*i

K-MEANS



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Figure 8.3. Using the K-means algorithm to find three clusters in sample data.

http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

K-MEANS

Assumptions:

- Clusters are spherical
- Clusters are well separated
- Clusters are of similar volumes
- Clusters have similar number of points

http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

K-MEANS

Q: How do you measure distance?
Q: How do you compute K?

http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

K-MEANS

Q: How do you measure distance?

A: Euclidean, many other distance metrics

Q: How do you compute K?

A: Problem specification, Trial and Error, Hierarchical Clustering, etc

http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

HIERARCHICAL CLUSTERING

Agglomerative Clustering

- Initially every point is its own cluster
- Merge the two nearest clusters (min, max, average, mean distance, etc)
- Repeat until all one cluster
- Output: Dendrogram

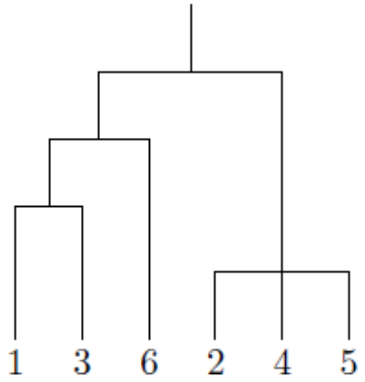


Figure 2: An example dendrogram