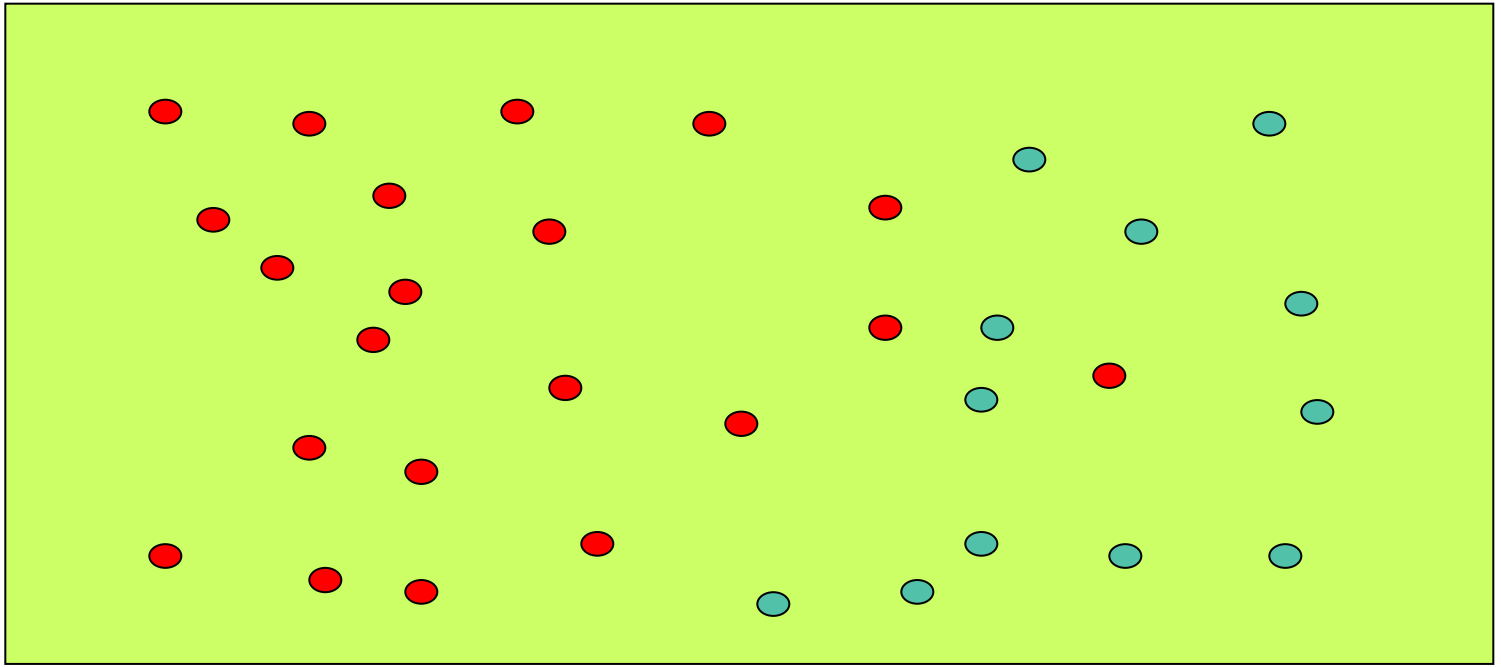
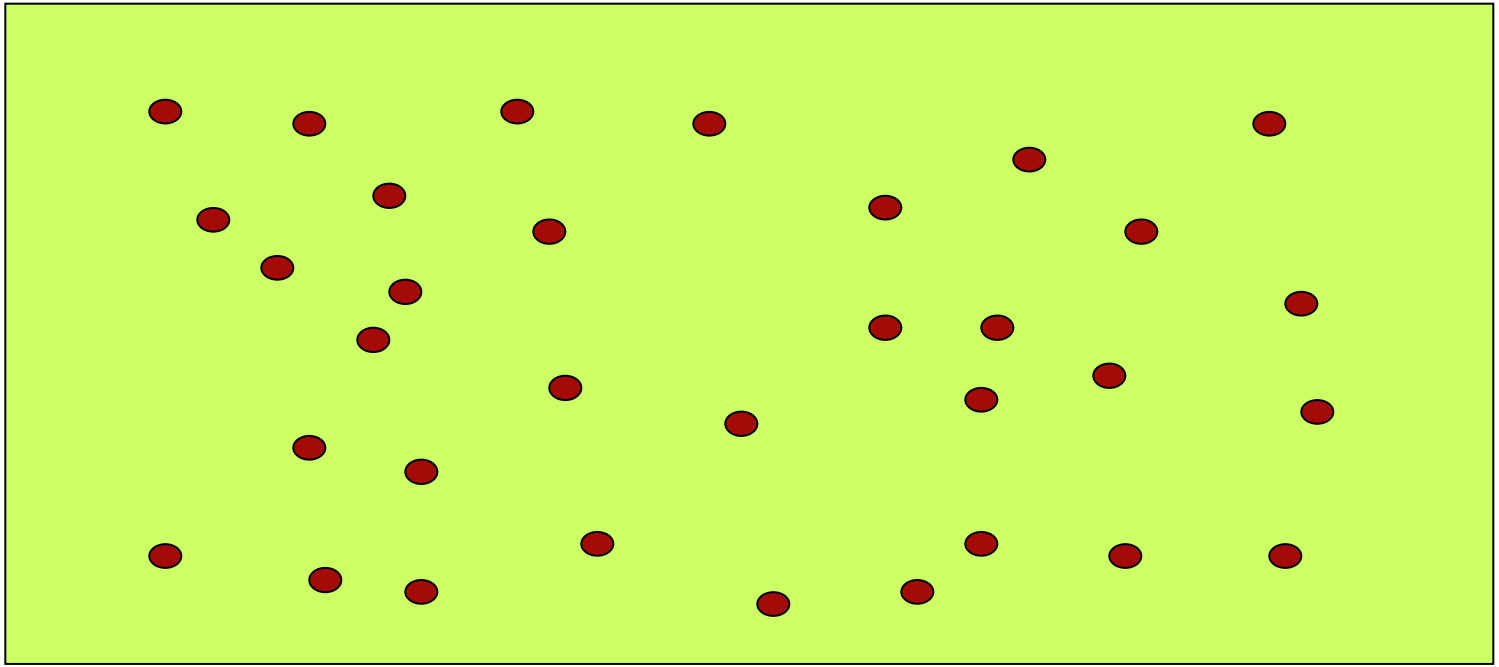
Cluster Analysis

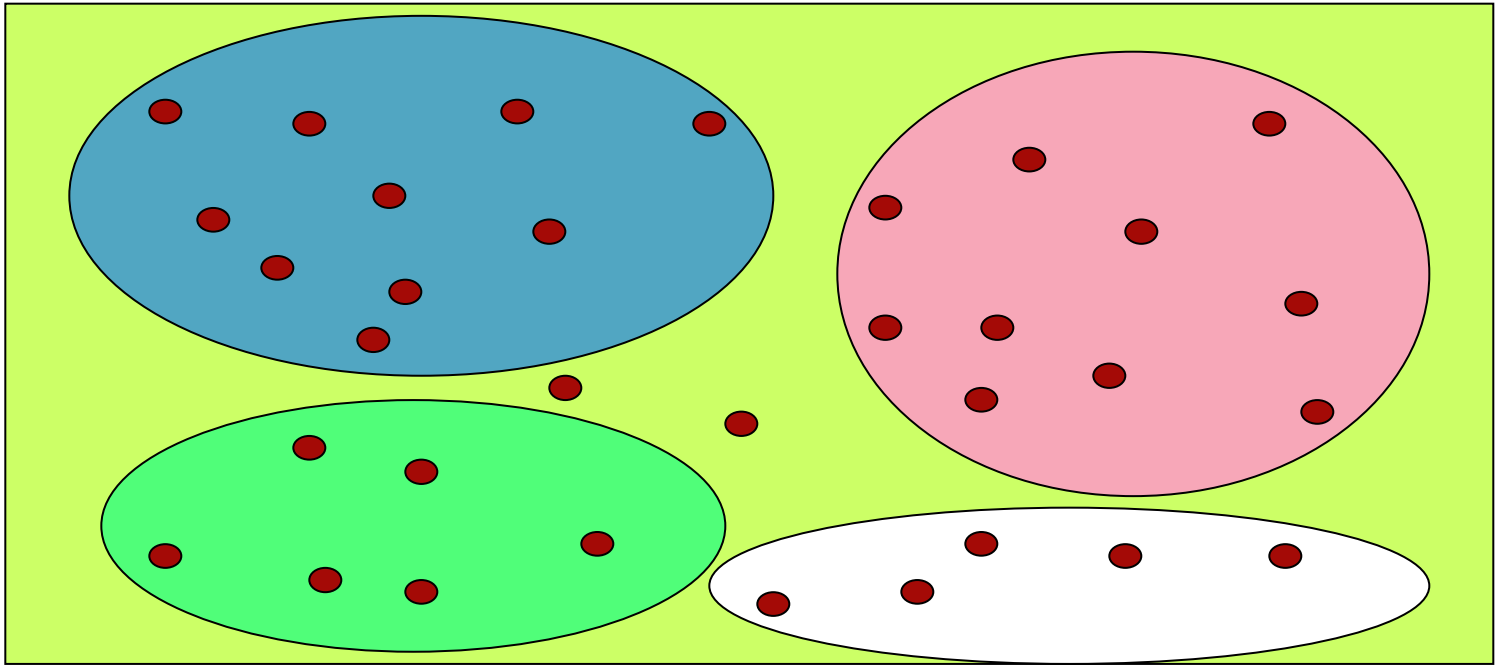
Labeled Training Data



Unlabelled Training Data



Possible Clusters



Cluster Analysis

- Discover similarity in data
  - Unsupervised learning
- Partition data into subsets (clusters) such that
  - Similarity within a sub-set is high — Intra-cluster
  - Similarity across sub-sets is low — Inter-cluster
- Tool for analysis of data
  - Better understanding of data
  - Visualization
- Pre-processing step
  - Numerosity reduction
  - Outlier removal

Applications

- Marketing
  - Customer Segmentation
- Documents
  - Topic detection
  - Information retrieval
- Image Processing
  - Segmentation
- Manufacturing
  - Shop Floor Management
    - Schedule similar jobs on the same machine; minimize retooling
- Security
  - Anomaly detection

Problem Setting

Some Definitions

The centroid of a cluster is the mean of the data points in the cluster.

$$\mu_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j$$

The radius of a cluster is the maximum distance from a data point to the centroid.

$$r_i = \max_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2$$

The diameter of a cluster is the maximum pairwise distance among data points.

$$d_i = \max_{x_p, x_q \in C_i} \left\| x_p - x_q \right\|^2$$

Clustering Approaches

- Partitional methods
  - Search among different partitions of the data
  - Methods: k-means, k-medioids, PAM, CLARANS

Clustering Approaches

- Hierarchical methods
  - Form a tree of clusters
    - Root contains all nodes, leaves are individual data points
  - Methods: AGNES, DIANA, BIRCH, CHAMELEON, ROCK

Clustering Approaches

- Density-based methods
  - Use notions of connectivity and reachability
  - Methods: DBScan, OPTICS, DenClue

- Grid-based approaches
  - Use inherent discretization

Clustering Approaches cont.

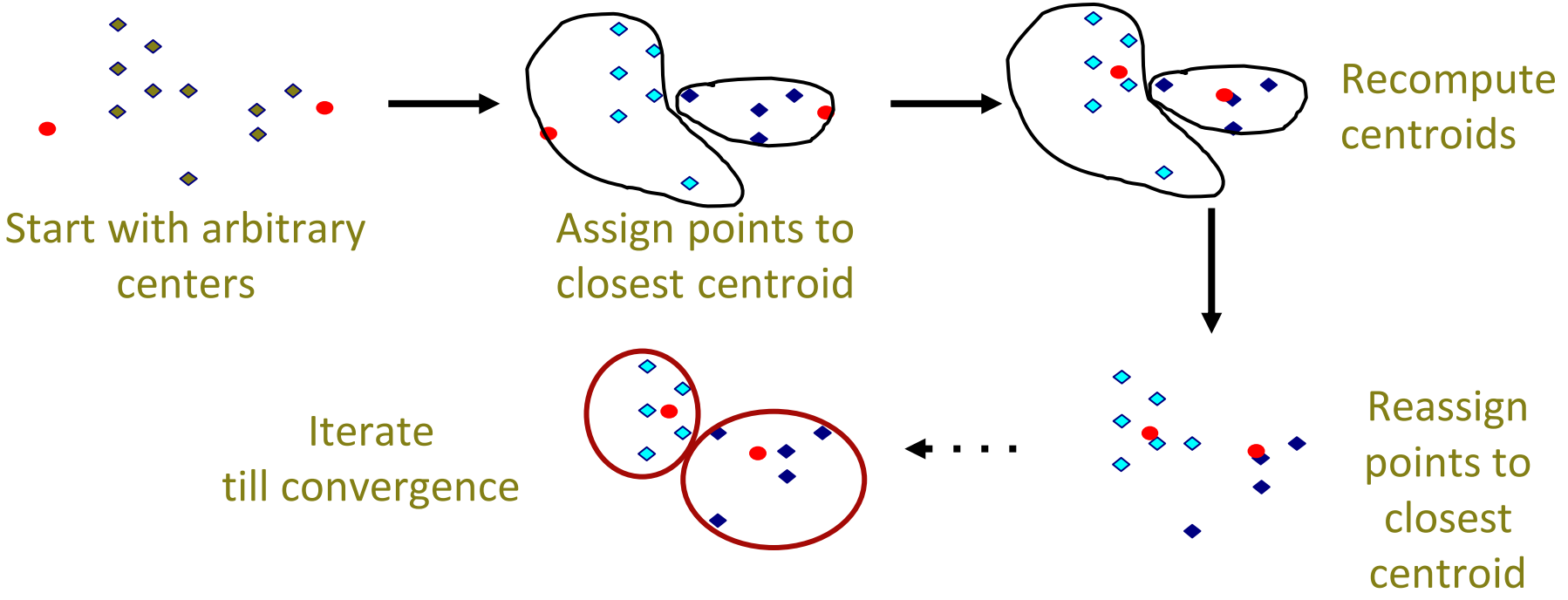
- Graph-based approaches
  - Use graph theoretic concepts
  - Spectral theory
- Constraint-based clustering
  - User specified constraints

K-means

- Most widely used!
- Simple to understand and use
- Reasonably fast
  - For small data sets
- Used to get a quick estimate of the *lay of the land*.
- Available in most data mining tools

K-means

K-means Illustrated



K-means algorithm

To cluster data into *k* groups:
(*k* is predefined)

1. Choose *k* cluster centers
   - e.g. at random
2. Assign instances to clusters
   - based on distance to cluster centers
3. Compute *centroids* of clusters
4. Make centroids as cluster centers. Go to step 2
   - until convergence

K-means clustering (in one slide!)

Problem: Find K cluster centers that minimize the sum of squared distance of each data point to the nearest cluster center.

Algorithm: Starting with K means initialized in some way, iterate until convergence:

a) ***[Centers -> Clusters] Assign*** each data point to the nearest mean.
b) ***[Clusters -> Centers] Update*** means to sample means of data points they are responsible for.

K-means clustering problem and notation

Input: N objects: $\{ x^{(n)} \} = x^{(1)}, x^{(2)}, \ldots, x^{(N)}$

Output: K means: $\{ m^{(k)} \} = m^{(1)}, m^{(2)}, \ldots, m^{(K)}$

(Squared) Distance measure: $d(x, m) = \Sigma_{i=1}^{I} (x_i - m_i)^2$, where I is dimensionality of data (2 in our examples).

Problem: Find K *means* that minimize $\Sigma_{n=1}^{N} d(x^{(n)}, m^{(k^*(n))})$, where

$k^*(n) = \text{argmin}_k \, d(x^{(n)}, m^{(k)})$

K-means clust. pseudocode: Lloyd's heuristic

Initialization: Set K means $\{m^{(k)}\}$ to random values.

Assignment: $r_k^{(n)} = 1$ if $m^{(k)}$ is the closest mean to datapoint $x^{(n)}$

0 otherwise

$$\mathbf{m}^{(k)} = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{R^{(k)}}$$

where $R^{(k)}$ is the total responsibility of mean k,

$$R^{(k)} = \sum_n r_k^{(n)}.$$

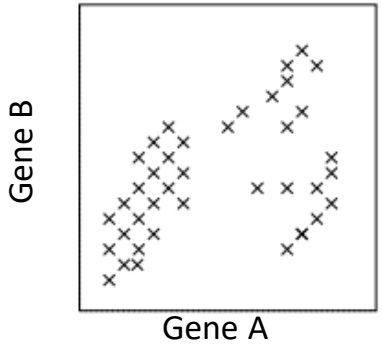
[DJM]

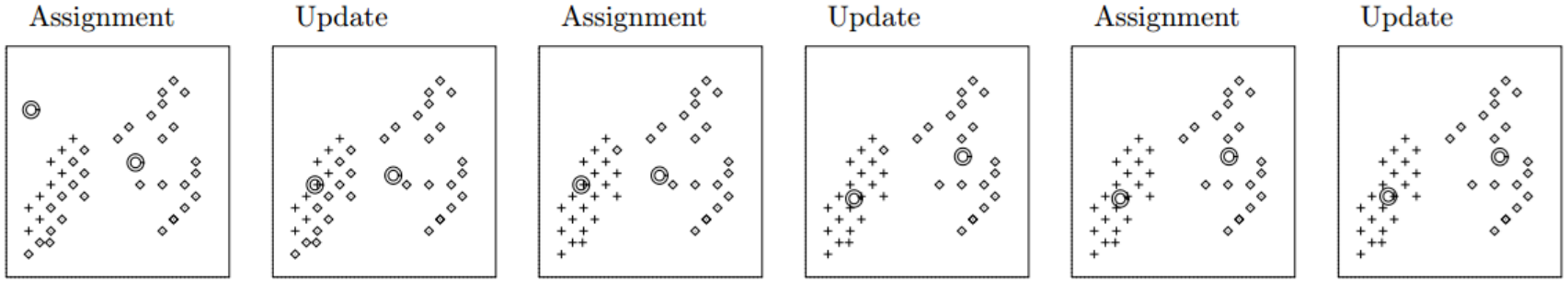
Food for thought – Bonus questions

- Prove that (hard) k-means yields clusters that are each convex sets!

- Prove that k-means is efficiently (polynomial-time) solvable when all datapoints lie on a real line (i.e., finding k centers that minimize k-means cost function when all datapoints are 1D).

Example run (K=2)



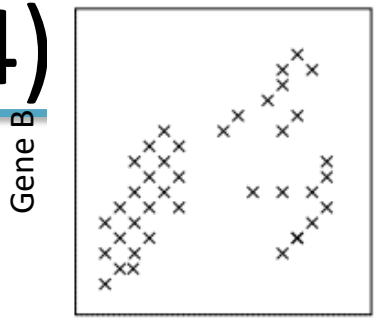


[DJM]

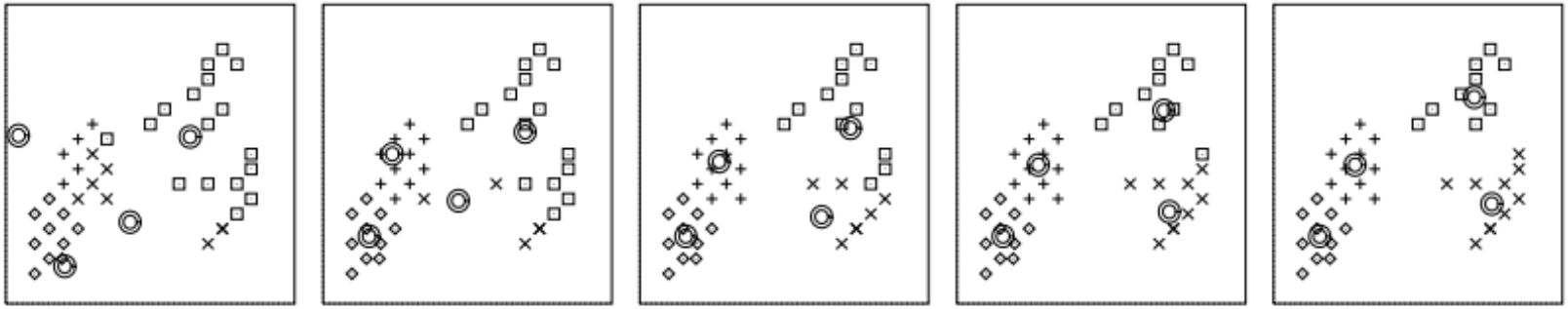
Let’s try K=4 clusters

with two different starting points!

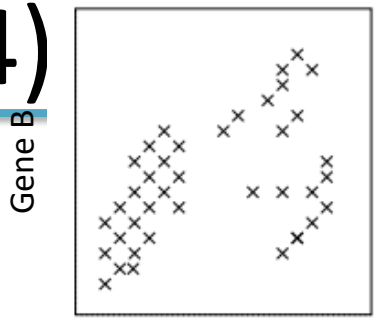
Example run1 (K=4)



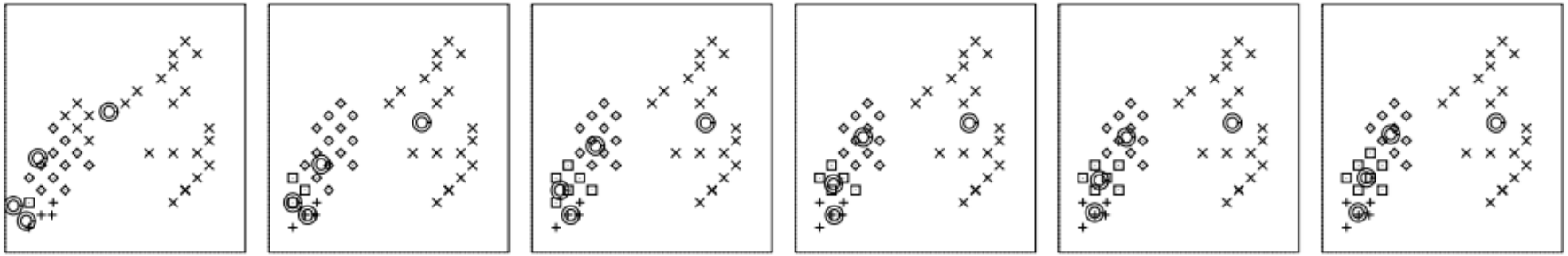
Run 1



Example run2 (K=4)



Run 2



[DJM]

Analysis of algorithm – Questions

Problem: Find K cluster centers that minimize the sum of squared distance of each data point to the nearest cluster center.

1) Are the means (centroids) the best (and **only**) cluster centers for a given partition?

2) Does the algorithm converge always for any dataset?

3) Does the algorithm actually (globally) minimize the sum of squared distances?

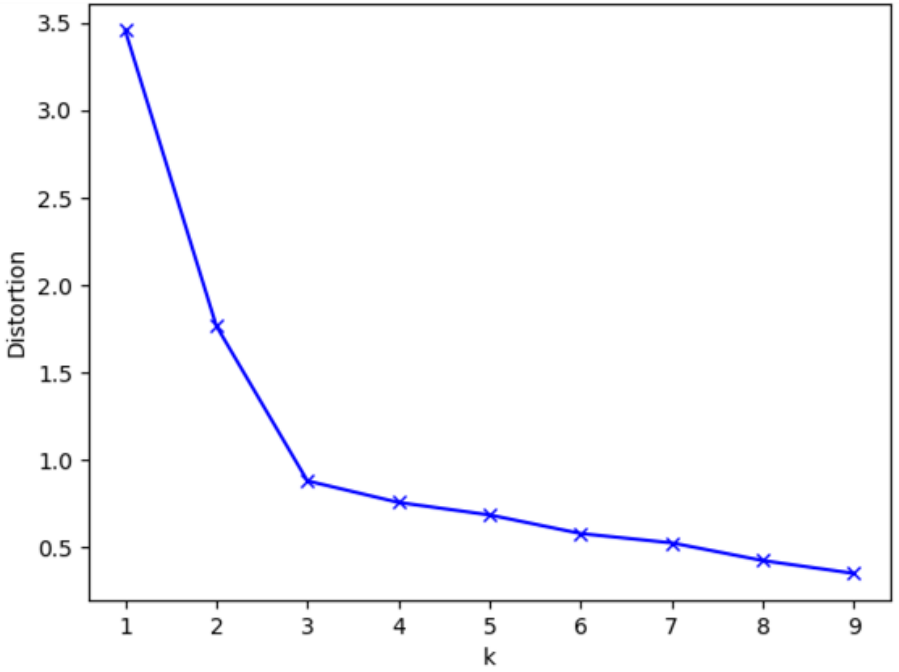
K-means critique

- Local optima
- Works with numeric data only
  - Centroids not defined for other types
  - Use actual data points: medoids
- Does not scale well to large data sets
  - Sampling
- Pre-specify number of clusters

The Knee Method

Finding the number of clusters

- Compute the total sum-of-squares error, for various values of k
- The location of a ‘knee’ is an indicator of the appropriate number of clusters
- The knee gives a small enough value of k with low error, past which we get diminishing returns
- As we increase k, the error will decrease, reaching zero when the number of clusters is equal to the number of datapoints.



K-means pitfalls: solutions?

-
- How about points in the border between two means? THINK FUZZY.

- What is a better alternative for distance $d(.,.)$? THINK ELLIPSOIDAL.
  - Instead of giving equal weight to all dimensions in a spherical fashion as in squared Euclidean distance, why not choose a distance that adapts to the width and breadth of each cluster?

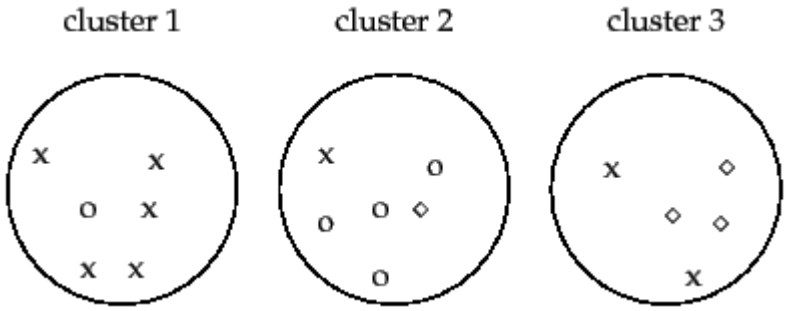
[DJM]

Quality of Clusters

- With labeled data
  - Purity, Entropy, ...
- With ground truth clustering
  - RAND Index, ...
- No information
  - Diameter, Radius, Dunn Index, Silhouette Score,...

Purity

- Each cluster is assigned to the class that is most frequent in the cluster
- Purity is measured by counting the number of correctly assigned datapoints and dividing by N
- A perfect clustering has a purity of 1, a bad clustering would have purity close to 0.



Let $\Xi = \{C_1, C_2, \cdots, C_k\}$ be the clustering.

Let $\Psi = \{\psi_1, \psi_2, \cdots, \psi_l\}$ be the classes,

with ψ_i being the data points belonging to class i.

Then, $purity(\Xi, \Psi) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{N_i} \max_j \left| C_i \cap \psi_j \right|$

Entropy

Let $\Xi=\{C_1,C_2,\cdots,C_k\}$ be the clustering and $\Psi=\{\psi_1,\psi_2,\cdots,\psi_l\}$ be the classes.

For each cluster the probability of a data point belonging to class j is given by:

$$P_{ij}=\left|C_i\cap\psi_j\right|/N_i$$

The entropy of the cluster is given by

$$e(C_i)=\sum_{j=1}^{l}-P_{ij}\log P_{ij}$$

The total entropy of the clustering is given by a weighted sum:

$$Entropy(\Xi,\Psi)=\sum_{i=1}^{k}\frac{N_i}{N}e(C_i)$$

Rand Index

Let $\Xi=\{C_1,C_2,\cdots,C_{k1}\}$ be the clustering and $\Omega=\{\omega_1,\omega_2,\cdots,\omega_{k2}\}$ be the reference clustering.

Ω is also referred to as the ground truth clusters.

Let A be the number of pairs of points that are in the same cluster in Ξ and in the same cluster in Ω.

Let B be the number of pairs of points that are in different clusters in Ξ and in different clusters in Ω.

Let P be the number of pairs of points that are in the same cluster in Ξ and but in different clusters in Ω.

Let Q be the number of pairs of points that are in different clusters in Ξ and but in the same cluster in Ω.

Then the Rand Index of the clustering is given by:

$$RandIndex(\Xi,\Omega)=\frac{A+B}{A+B+P+Q}=\frac{A+B}{N(N-1)/2}$$

The Rand Index can be thought of as the probability that the clustering and the ground truth would agree on a randomly chosen pair of points.

Silhouette Score

For a data point $x_j \in C_i$

Let $a(x_j)$ be the average distance of x_j to the other points in C_i.

Let $b(x_j) = \min_{k \neq i}$ average of the distance of x_j to each of the data points in C_k.

This is the average distance to the *neighbour cluster*.

$$Silhouette(x_j) = \frac{b(x_j) - a(x_j)}{\max\{a(x_j), b(x_j)\}}$$

This score lies between -1 and +1.

If the score for a data point is close to -1, then it is better off being assigned to the neighbour cluster.

The silhouette score of the clustering is the mean of the scores of all the points.

The closer it is to +1 the better the clustering.

Silhouette Score

For a data point $x_j \in C_i$

Let $a(x_j)$ be the average distance of x_j to the other points in C_i.

Let $b(x_j) = \min_{k \neq i}$ average of the distance of x_j to each of the data points in C_k

This is the average distance to the *neighbour cluster*.

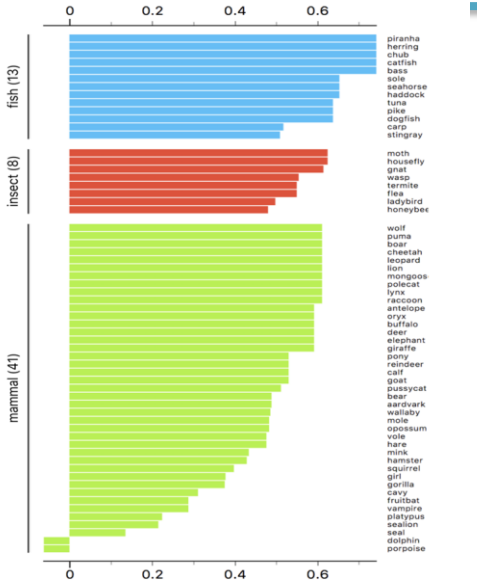
$$Silhouette(x_j) = \frac{b(x_j) - a(x_j)}{\max\{a(x_j), b(x_j)\}}$$

This score lies between -1 and +1.

If the score for a data point is close to -1, then it is better off being assigned to the neighbour cluster.

The silhouette score of the clustering is the mean of the scores of all the points.

The closer it is to +1 the better the clustering.



Dunn Index

Let $\delta(C_i, C_j)$ denote the inter cluster distance between C_i and C_j.

Let $d(C_i)$ denote the diameter of cluster C_i

$$DunnIndex(\Psi) = \frac{\min_{1 \le i < j \le k} \delta(C_i, C_j)}{\max_i d(C_i)}$$

Measure of compactness of clusters, but suffers from outlier probelems.

If one cluster has a high diameter then it brings down the quality of the overall clustering.

Can also be used to determine the number of clusters.

Hierarchical Clustering

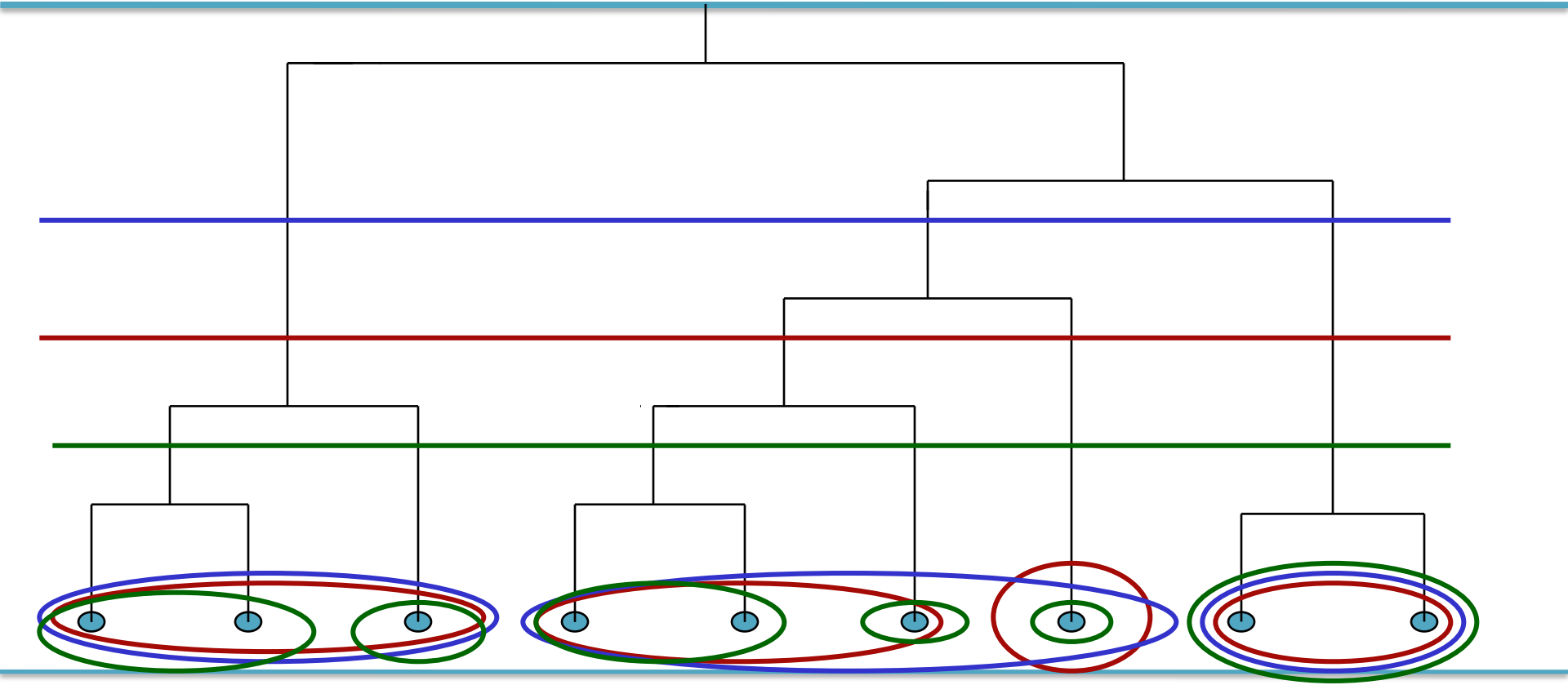
- Build trees of clusters
  - Dendrograms
- Agglomerative Clustering
  - Start with individual points
  - Repeatedly merge closest clusters based on some criterion
  - Most popular
- Divisive Clustering
  - Start with one cluster
  - Divide based on some criterion

Dendrogram: How the Clusters are Merged

- Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

Dendrogram: How the Clusters are Merged



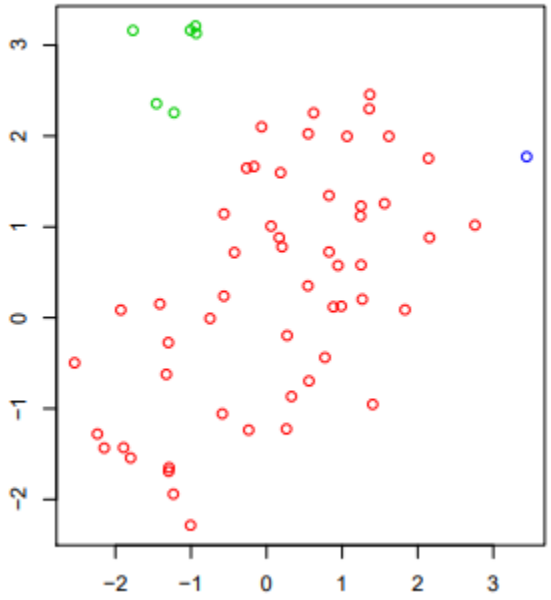
Hierarchical Agglomerative Clustering

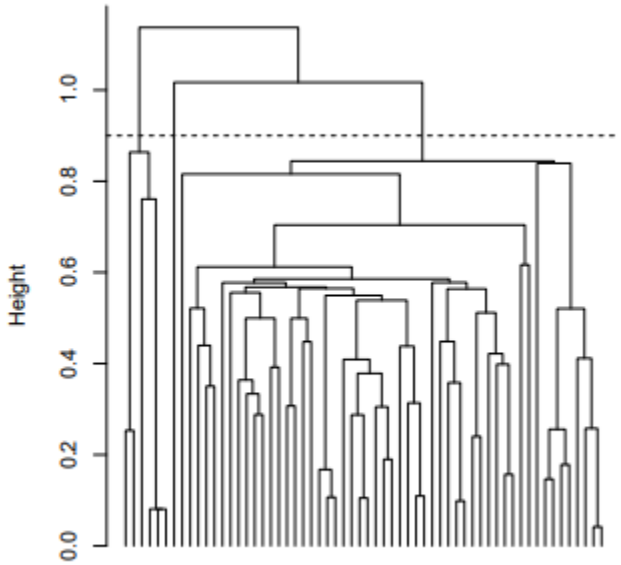
- Number of clusters not pre-specified
  - Some stopping criterion is required
- Computationally expensive
  - Scaling techniques
    - Sampling
    - Clustering features

Inter Cluster Distance

- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $\text{dist}\left(C_i, C_j\right) = \min_{x_p \in C_i, x_q \in C_j} \text{dist}\left(x_p, x_q\right)$
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., $\text{dist}\left(C_i, C_j\right) = \max_{x_p \in C_i, x_q \in C_j} \text{dist}\left(x_p, x_q\right)$
- Average: avg distance between an element in one cluster and an element in the other, i.e., $\text{dist}\left(C_i, C_j\right) = \frac{1}{N_i N_j} \sum_{x_p \in C_i, x_q \in C_j} \text{dist}\left(x_p, x_q\right)$
- Centroid: distance between the centroids of two clusters, i.e., $\text{dist}\left(C_i, C_j\right) = \text{dist}\left(\mu_i, \mu_j\right)$
- Medoid: distance between the medoids of two clusters, $\text{dist}\left(C_i, C_j\right) = \text{dist}\left(\text{medoid}\left(C_i\right), \text{medoid}\left(C_j\right)\right)$
  - Medoid: one chosen, centrally located object in the cluster

Single Link

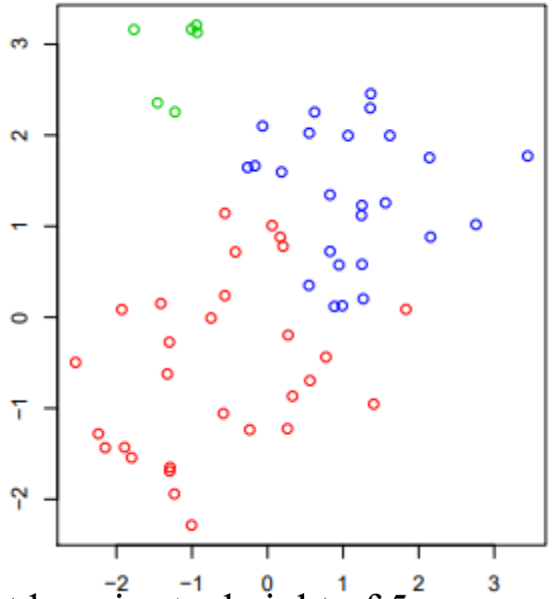


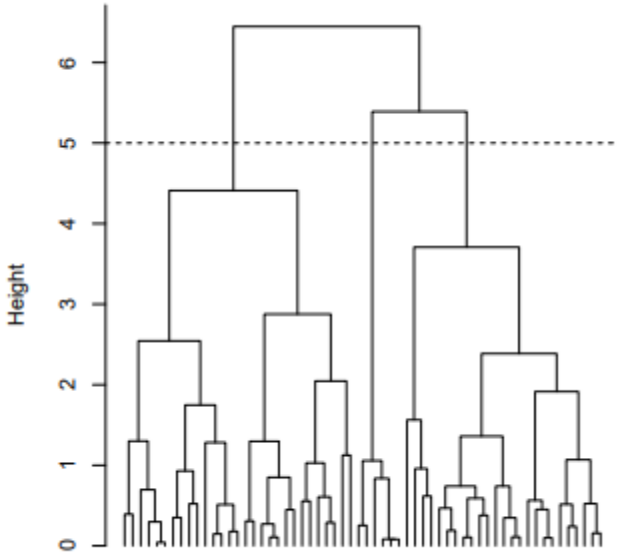


The cut here is at a height of 0.9.

For each point x_i, there is a point x_j at a distance less than 0.9.

Complete Link

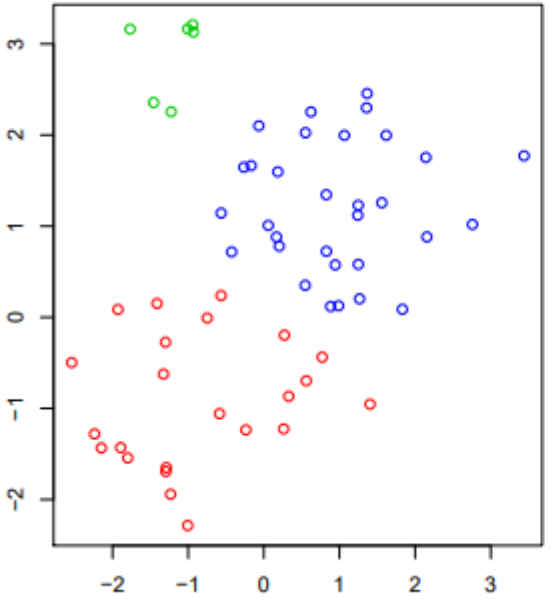


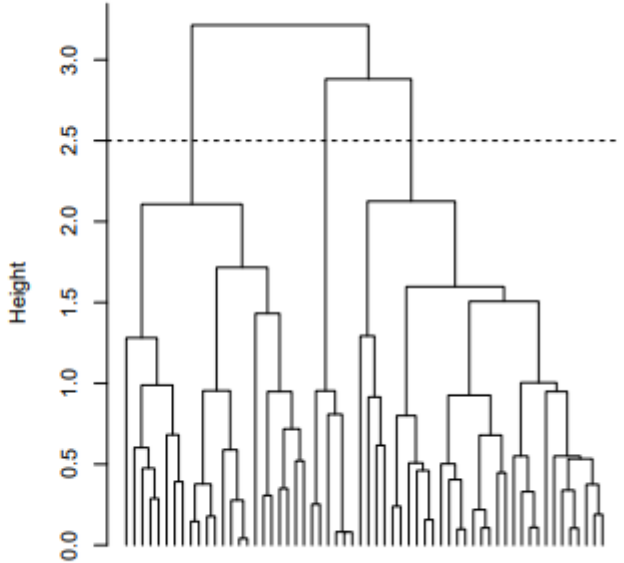


The cut here is at a height of 5.

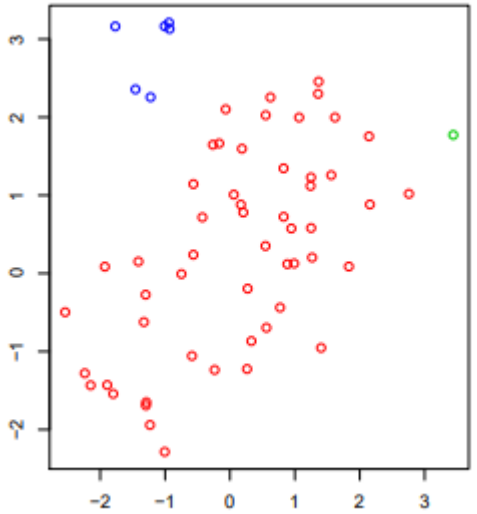
For each point x_i, every other point in its cluster is within a distance less than 5.

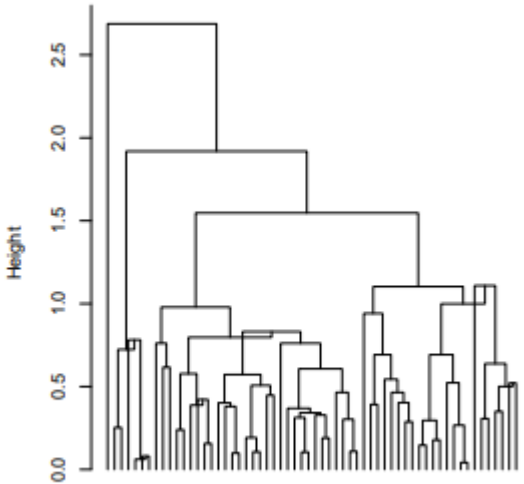
Average Link





Centroid Link

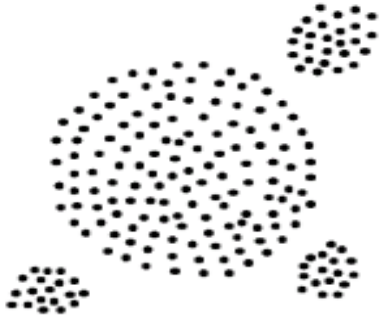




Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

Density-based Clusters



database 1

database 2

database 3

DB Scan

- *Eps* and *minPts*
- *Core Point*
- *Density Reachable*
- *Density Connected*
- *Border Point*

DBSCAN: The Algorithm

- Arbitrary select a point *p*
- Retrieve all points density-reachable from *p* w.r.t. *Eps* and *minPts*
- If *p* is a core point, a cluster is formed
- If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database
- Continue the process until all of the points have been processed

DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

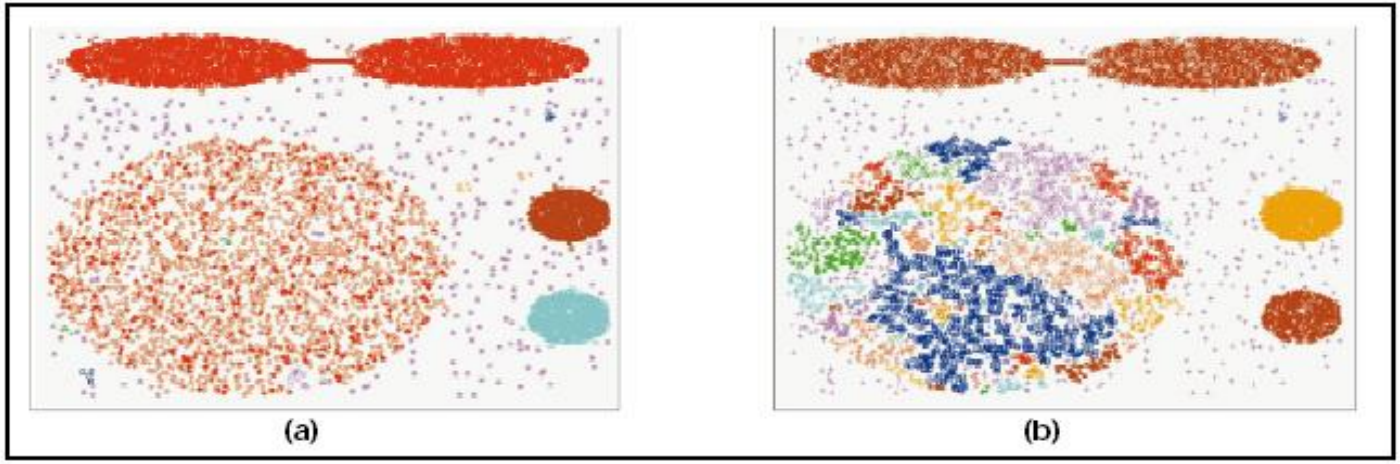
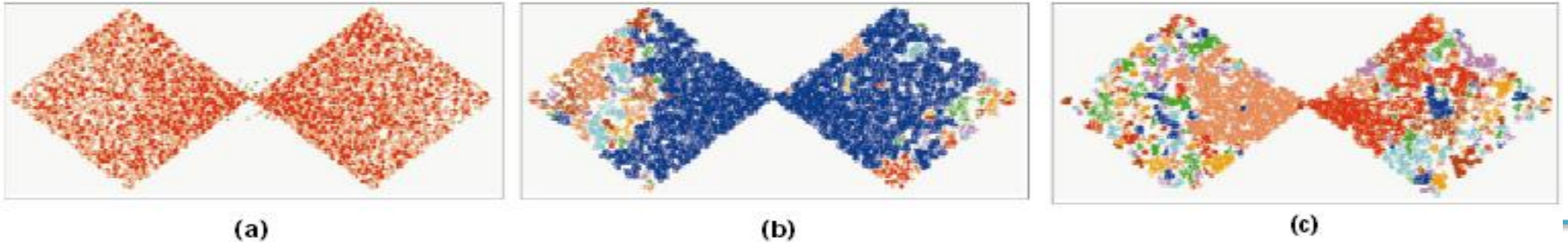


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



Spectral Clustering

- Given a set of data points A, a similarity matrix S may be defined where S_{ij} represents a measure of the similarity between points i and j ($i, j \in A$)
- Spectral clustering makes use of the spectrum of the similarity matrix of the data to perform dimensionality reduction for clustering in fewer dimensions
  - In functional analysis, the spectrum of a bounded operator is a generalization of eigenvalues for matrices
    - A complex number λ is said to be in the spectrum of a bounded linear operator T if $\lambda I - T$ is not invertible, where I is the identity operator

Outlier Detection

- Outlier: not part of the data distribution
- Pre-processing
  - Noise Removal
- Objective
  - Anomaly detection
- Density based methods more suited

Numerosity Reduction

- Form tight clusters
- Use representative points
  - Centroids
  - Medoids
  - Multiple points per cluster
- Key idea behind some scaling techniques

Summary

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- Quality of clustering results can be evaluated in various ways