slides originally by
Dr. Richard Burns,
modified by
Dr. Stephanie Schwartz

CLUSTERING

CSCI 452: Data Mining

Topic Outline

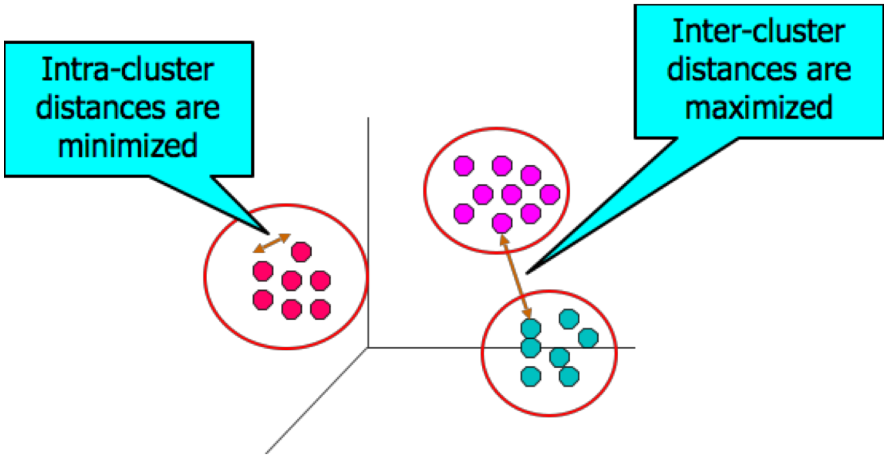
- Unsupervised Learning
- K-Means Clustering
- Hierarchical Clustering
- Cluster Evaluation

What is Cluster Analysis?

- Different than *prediction*…
- Dividing data into groups (clusters) in some meaningful or useful way
- Clustering should capture “natural structure of the data”

What is Cluster Analysis?

- Finding groups of objects such that the objects within a group will be *similar* to one another and *different* from the objects in other groups



Applications for Clustering

- Starting point for data summarization
- Understanding
  - (human beings are good at dividing things into groups that share common characteristics)

Specific Applications

- Business
  - Businesses collect large amounts of information on current and potential customers.
  - Clustering to segment customers into a small number of groups, for additional analysis and marketing activities.

Clustering for Utility

- Efficiently Finding Nearest Neighbors
  - Alternative to computing the pairwise distance between all points
  - If we already have clusters…
    - … then we usually only need to know the pairwise distance for only the objects within the cluster of interest
    - Considerably less computations

Notion of a Cluster can be Ambiguous

- The notion of a cluster may not be well defined.

How many clusters?

How many clusters?

By the human visual system, it looks like two clusters.

Two Clusters

But it really depends on the characteristics of the data.
These clusterings may not be unreasonable:

Six Clusters

Four Clusters

Relation of Clustering to Classification

- Clustering can be regarded as a form of classification
  - Creating a labeling of objects with cluster (class) labels
  - But…these labels are derived exclusively from the data.
  - Cluster analysis is sometimes referred to as <u>unsupervised classification</u>
    - No model from training data with class labels

Iris Example

- With Decision Trees (supervised classification):
  - "Training set" has class labels:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.9	3.0	1.4	0.2	setosa
4.6	3.1	1.5	0.2	setosa
6.7	3.1	4.4	1.4	versicolor
6.4	2.8	5.6	2.2	virginica

- With Clustering (unsupervised classification):
  - Only data

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
4.9	3.0	1.4	0.2
4.6	3.1	1.5	0.2
6.7	3.1	4.4	1.4
6.4	2.8	5.6	2.2

Types of Clusterings

- Partitional vs. Hierarchical
  - Partitional Clustering: A division of data into non-overlapping clusters, such that each data object is in exactly one subset
  - Hierarchical Clustering: A set of nested clusters organized as a hierarchical tree
    - Each node (cluster) is union of its children (subclusters)
    - Root of tree: cluster containing all data objects
    - Leaves of tree: singleton clusters

Types of Clusterings

- Complete vs. Partial
  - Complete Clustering: Every object is assigned to a cluster
  - Partial Clustering: Not every object needs to be assigned
    - *Motivation:* some objects in a dataset may not belong to well-defined groups
      - Noise, outliers, or simply "uninteresting background" data

Types of Clusterings

- Exclusive vs. Non-exclusive
  - Exclusive Clustering: Assignment is to *one* cluster
  - Non-Exclusive Clustering: Data objects may belong to multiple clusters
    - *Motivation:* multiclass situations
      - *Example*: "student employee"

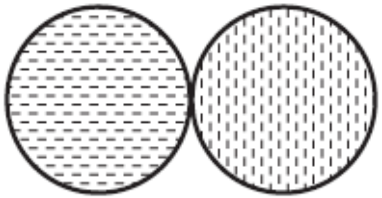
Types of Clusters

- Well-Separated Clusters
  - any point in a cluster is closer to every other point in the cluster than to any point not in the cluster

(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

Types of Clusters

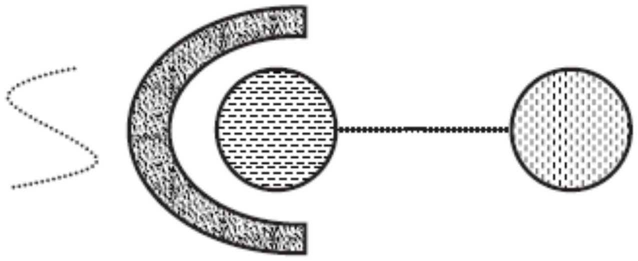
- Center-based Clusters
  - an object in a cluster is closer to the *center* of a cluster than to the *center* of any other cluster
  - Center of a cluster ("the most central point"):
    1. Centroid: the mean of all the points in the cluster *(usually for continuous attributes)*
    2. Medoid: the most "representative" point of a cluster *(usually for categorical attributes)*



(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

Types of Clusters

- Contiguous Clusters
  - a point in a cluster is closer to one or more other points in the cluster than to any point not in the cluster



(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.

Clustering Algorithms

1. K-means
2. Hierarchical

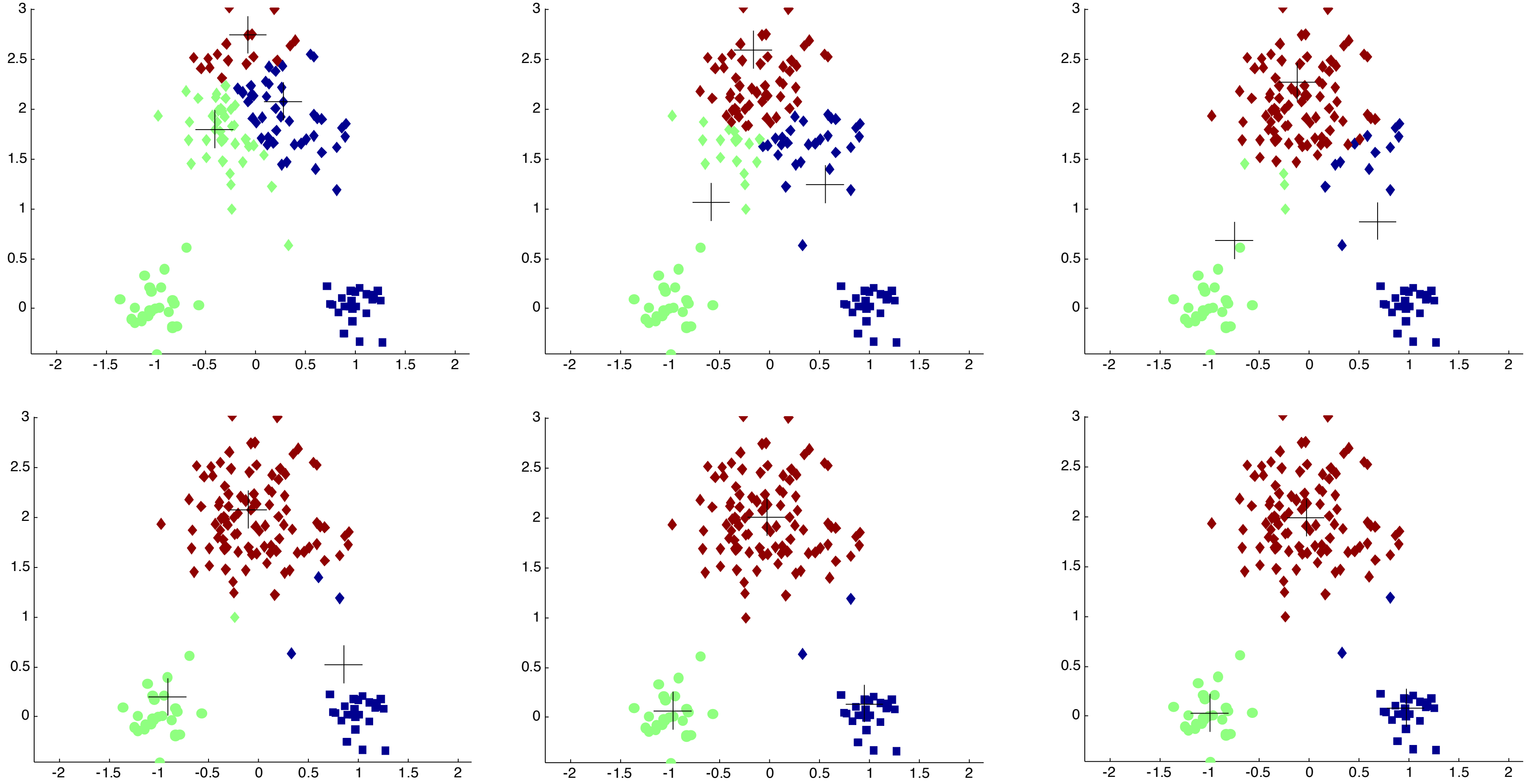
K-means Clustering

Prototype-Based Clustering:

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

- *k* is chosen by data analyst
- k = *number of clusters*

Using K-means to find three clusters in sample data: *first 6 iterations are shown*

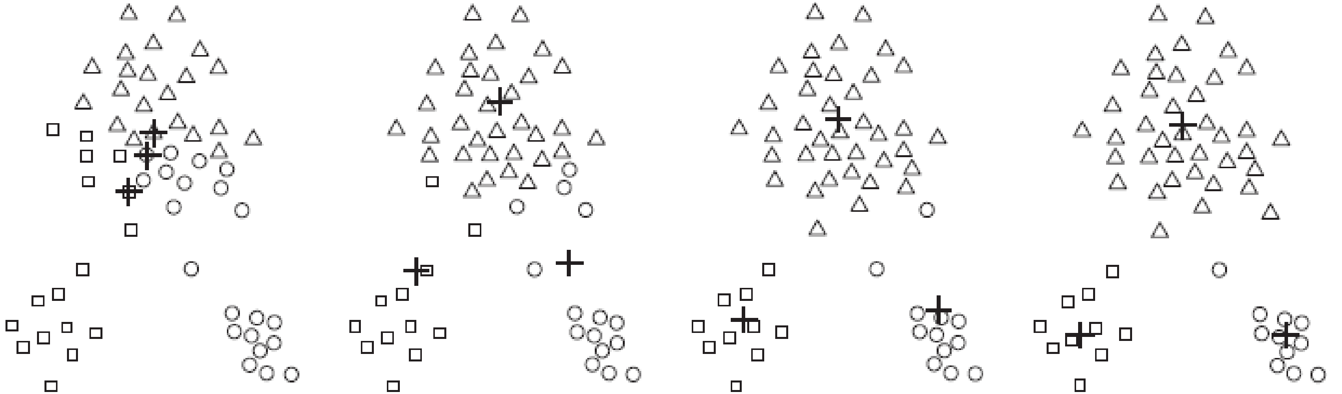


How to Choose Initial Centroids?

- One *strategy:* choose the *k* centroids at random

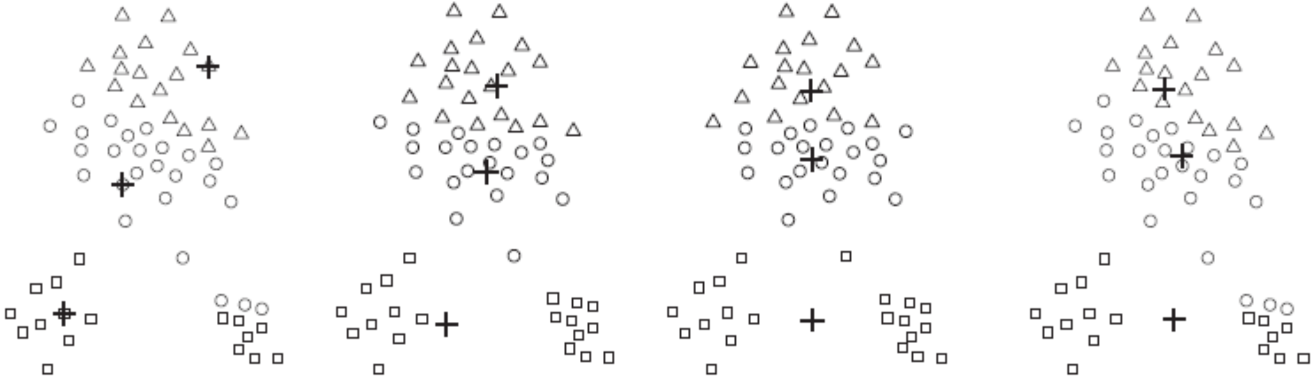
- Different runs of k-means on same data:
  - Will produce different iterations (because the starting clusters are different)
  - May produce different final clusters

Optimal Clustering:



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

Sub-Optimal Clustering:



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

When to Stop the Iterating Process?

- Most convergence happens in the first few iterations
- Sometimes the *termination condition* is:
  - "repeat until only 1% of the points change clusters"

Assigning Points to the Closest Centroid

- Decided by the analyst
  - Euclidean distance
  - Manhattan distance

Cluster Evaluation

- Determining the quality of a clustering
- <u>Sum of the Squared Error (SSE):</u>
  - Objective function
  - Calculation of the “error” of each data point:
    - Distance to the centroid
  - Total sum of the squared errors

Given two different clusterings, produced by two different runs of k-means, prefer the clustering with the smaller SSE.

- Clustering is a better representation

Solutions to Initial Centroids Problem

- Multiple runs
- Select more than k initial centroids
  - Then select the k most widely separated
    - Computational issues?
    - Bias toward centroid being affected by far-away outliers
- … and others

Other Potential Issues

- "basic" *k*-means may yield empty clusters
  - *One strategy:* assign to the empty cluster the point that is farthest away from any current centroid
    - Computationally expensive?
  - *Another strategy:* choose a point from the cluster with the highest SSE
    - Effect: will split the cluster and reduce the overall SSE

Other Potential Issues

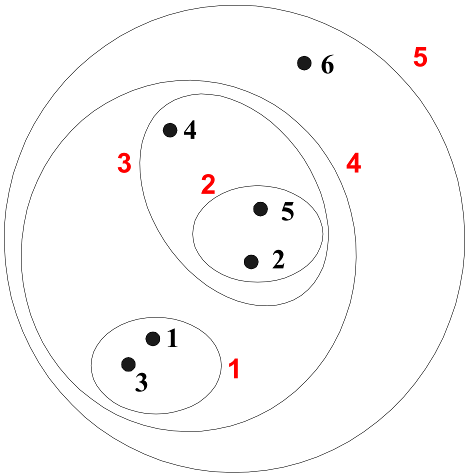
- Outliers can unduly influence the clusters that are found:
  - The resulting cluster centroids may not be as representative as they otherwise would be
  - Often useful to try to eliminate outliers beforehand

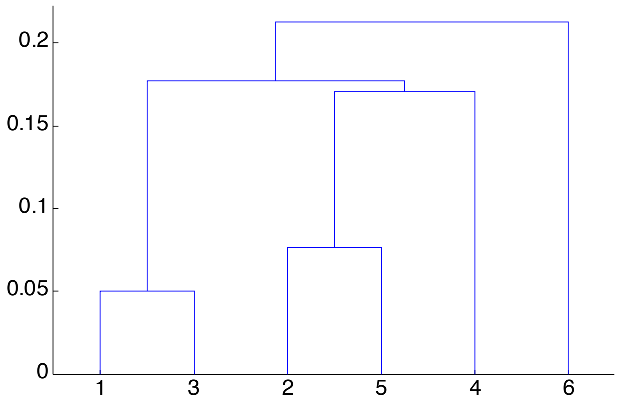
Clustering Algorithms

1. K-means
2. <u>Hierarchical</u>

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits





Hierarchical Clustering Approaches

1. Agglomerative: start with data points as *individual clusters (bottom-up)*
   - at each step merge the closest pair of clusters
   - *Definition of "cluster proximity" needed.*
2. Divisive: start with one all-inclusive cluster *(top-down)*
   - at each step split a cluster until only singleton clusters remain
   - Need to decide which cluster to split and how to do splitting

Why Hierarchical Clustering?

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level

- They may correspond to meaningful taxonomies
  - *Example:* biological sciences domain (e.g., animal kingdom, phylogeny reconstruction, …)

Agglomerative Clustering Algorithm

1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

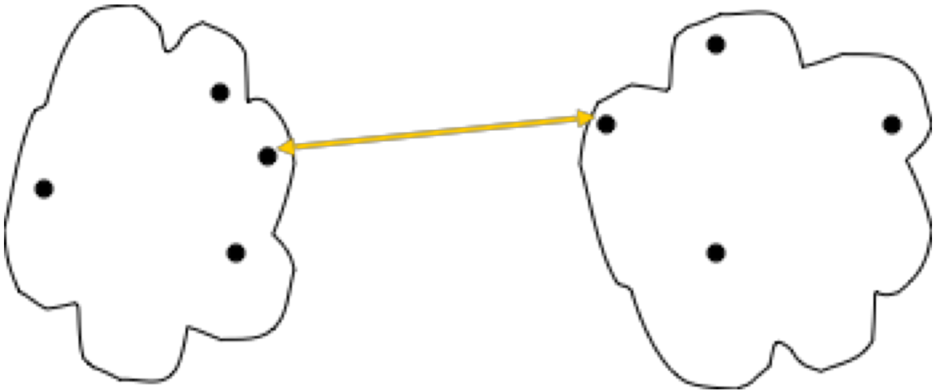
Originally, the distance between two points

Update with distance between two clusters.

- How to define?

Defining Proximity between Clusters

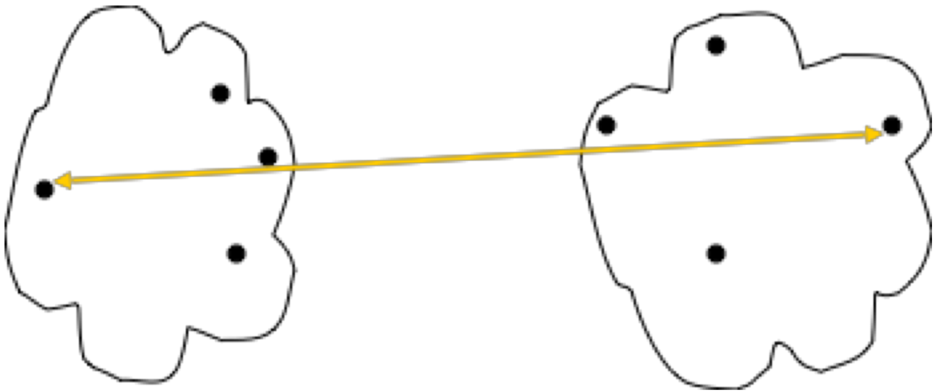
MIN (single-link)



Graph-Based view

Defining Proximity between Clusters

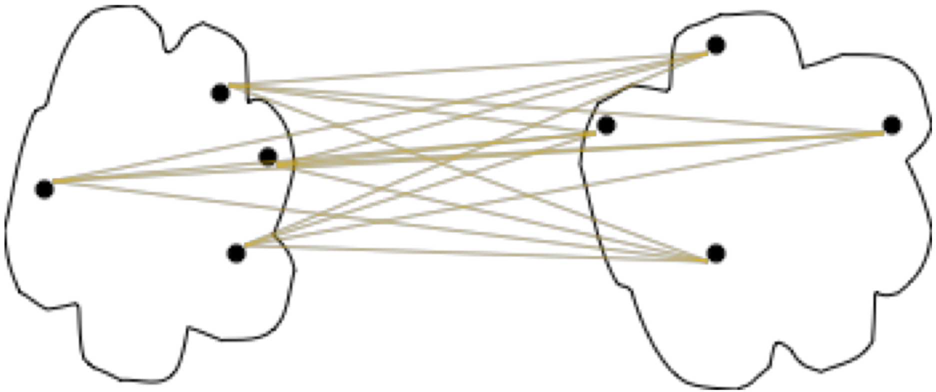
MAX (complete-link)



Graph-Based view

Defining Proximity between Clusters

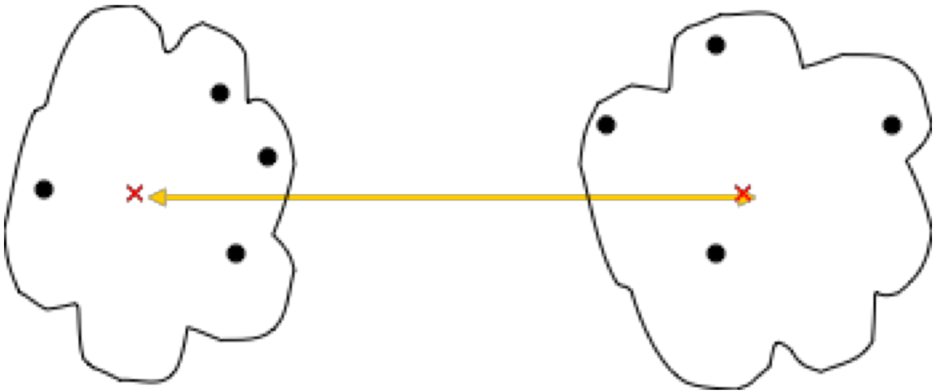
Group Average



Graph-Based view

Defining Proximity between Clusters

Distance between Centroids



Prototype-Based view

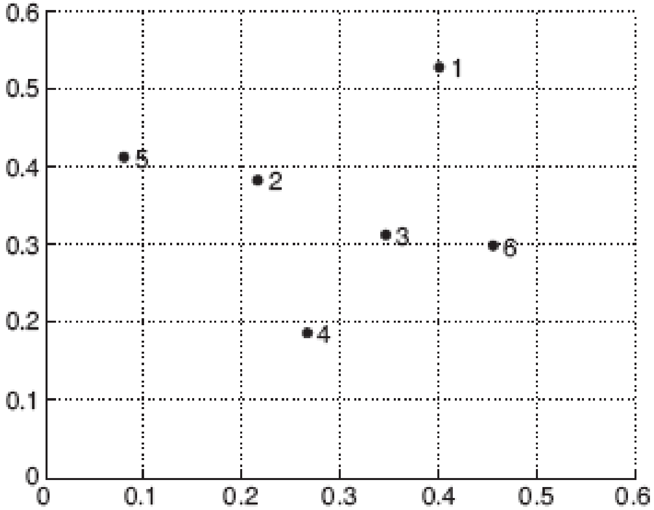
Choice of Proximity Measure Will Affect Results

- MIN
  - Can handle non-elliptical shapes
  - Sensitive to noise and outliers
- MAX
  - Less susceptible to noise and outliers
  - Tends to break large clusters
  - Biased towards globular clusters
- Group Average
  - Compromise between MIN and MAX

What's the first merge?

Quick Partial Example using MIN

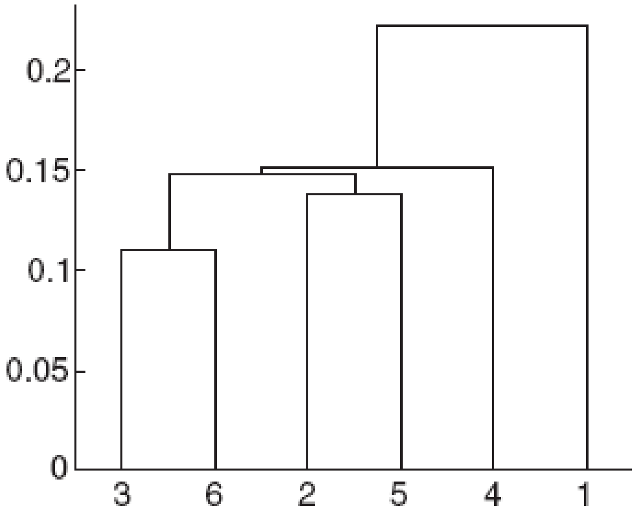
- Data points and Proximity Matrix:



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Total space complexity: $O(n^2)$

Quick Partial Example using MIN



Let's say we now have:

- Cluster {3,6}
- Cluster {2,5}
- Cluster {4}
- Cluster {1}

What's the next merge?

MIN({3,6},{2,5}) = 0.15
Is the next smallest value

Hierarchical Clustering: Time Complexity

- Space complexity: $O(n^2)$
- Time complexity:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

- $O(n^3)$
  - n steps (number of merges)
  - At each step: proximity matrix must be searched: n^2

Hierarchical Clustering

- No backtracking:
  - Once a decision is made to combine two clusters, it cannot be undone

More Advanced…

- Determining the Correct Number of Clusters
  - Try different numbers of clusters
  - Statistical measures used to help decide:
    1. SSE
    2. “Silhouette Coefficient”

References

- *Introduction to Data Mining*, 1st edition, Tan et al.
- *Data Mining and Business Analytics with R*, 1st edition, Ledolter