CS 412 Introduction to Machine Learning

# Agglomerative clustering

Instructor: Wei Tang

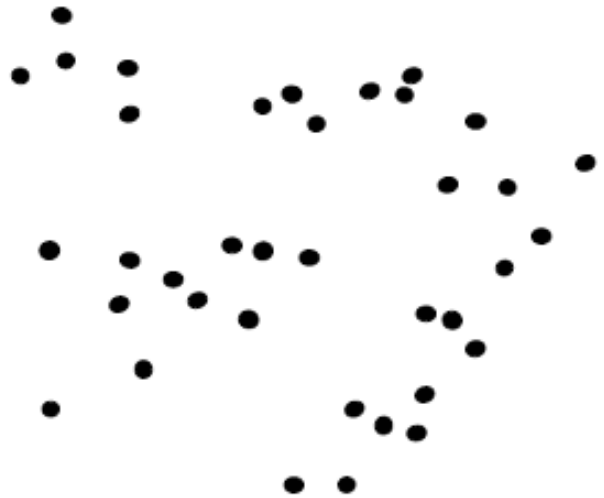
Department of Computer Science
University of Illinois at Chicago
Chicago IL 60607

https://tangw.people.uic.edu
tangw@uic.edu

# Announcements

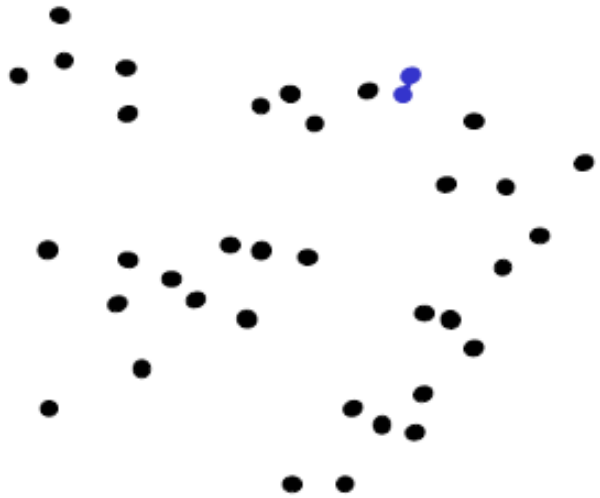
- Machine problem #4 available on Blackboard
  - Due on 11/17 (Wed)
  - Clustering
  - Last homework!

# Agglomerative clustering



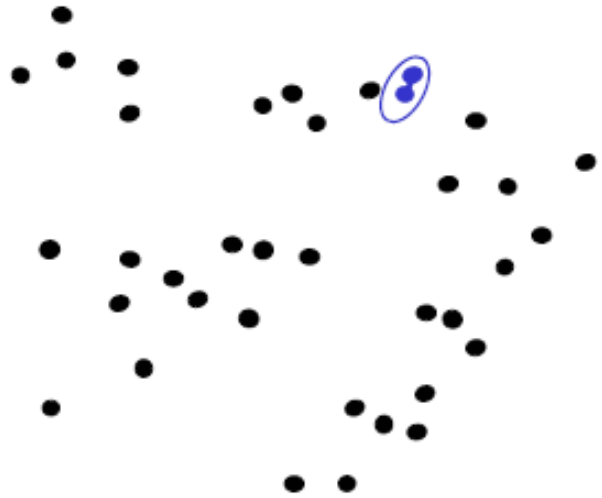
1. Say "Every point is its own cluster"

# Agglomerative clustering



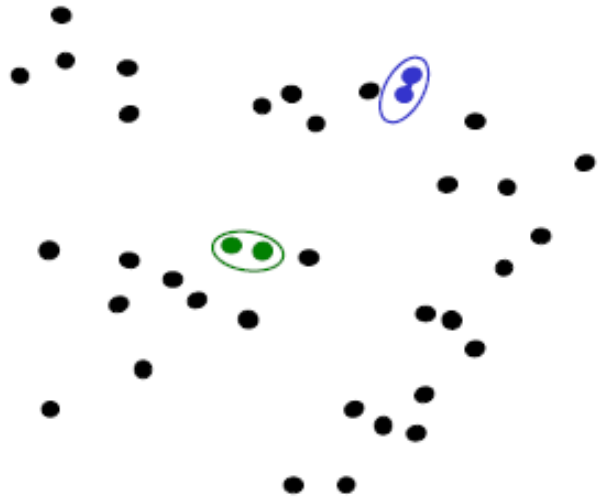
1. Say "Every point is its own cluster"
2. Find "most similar" pair of clusters

# Agglomerative clustering



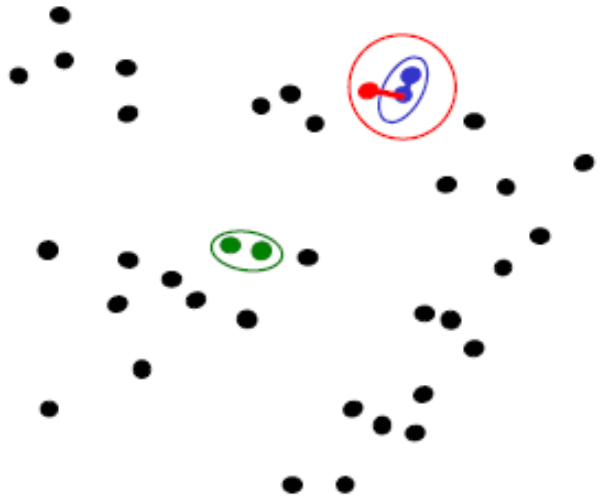
1. Say "Every point is its own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster

# Agglomerative clustering

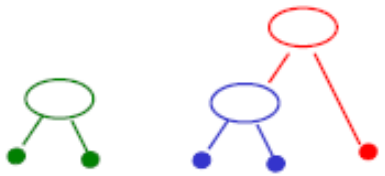


1. Say "Every point is its own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster
4. Repeat

# Agglomerative clustering



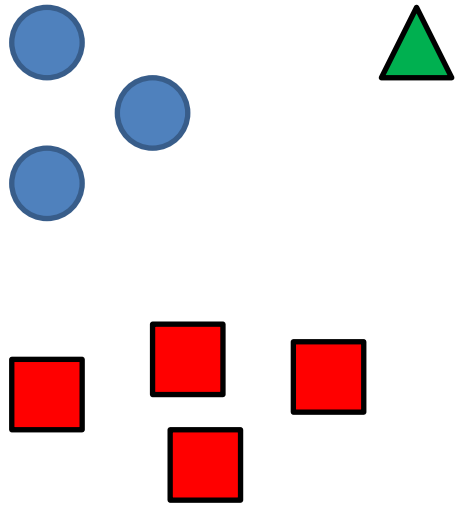
1. Say "Every point is its own cluster"
2. Find "most similar" pair of clusters
3. Merge it into a parent cluster
4. Repeat



# Agglomerative clustering

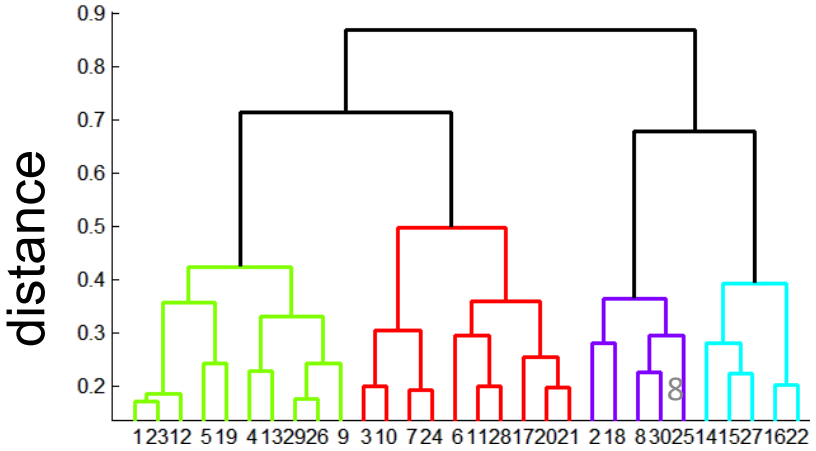
## How to define cluster similarity?

- Average distance between points, maximum distance, minimum distance
- Distance between means



## How many clusters?

- Clustering creates a tree
- Threshold based on max number of clusters or based on distance between merges



# Conclusions: Agglomerative Clustering

## Good

- Simple to implement, widespread application
- Clusters have adaptive shapes
- Provides a hierarchy of clusters

## Bad

- May have imbalanced clusters
- Still have to choose number of clusters or threshold