



# CMSC726 Spring 2006: Unsupervised Learning

readings:  
sources: course slides are based on material from a variety of sources, including **Tom Dietterich**, Carlos Guestrin, Terran Lane, **Rich Maclin**, Ray Mooney, **Andrew Moore**, Andrew Ng, Jude Shavlik, **Sebastian Thrun** and others.




## Unsupervised Learning

- Supervised learning: Data  $\langle x, y \rangle$
- Unsupervised Learning: Data  $x$




## Goals of Unsupervised Learning

- Clustering
- Visualization
- Density Estimation
- Outlier/Novelty Detection
- Data Compression



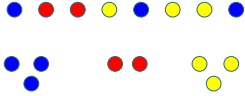

## Descriptive Models

- model presents the main features of the data, a global summary of the data
  - cluster analysis
  - density estimation



## Cluster Analysis

- decomposing or partitioning a data set into groups so that
  - the points in one group are similar to each other
  - and are as different as possible from the points in other groups

## General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document indexing by topic
  - Cluster Weblog data to discover groups of similar access patterns

### ● ● ● | Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with similar claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

### ● ● ● | Example

- households:  
location, income, number of children, rent/own, crime rate, number of cars
- The appropriate clustering will depend on goals:
  - minimize delivery time  $\Rightarrow$  cluster by location
  - others?

### ● ● ● | Clustering

- Decomposing or partitioning data into groups so that
  - the points in one group are **similar** to each other
  - and are as **different** as possible from the points in other groups
- Measure of **distance** is fundamental
- Explicit representation:
  - $D(x(i), x(j))$  for each  $x$
  - only feasible for small domains
- Measurement:
  - distance computed from features
  - we've already seen a number of different methods

### ● ● ● | Clustering

- **Huge** body of work
- (aka unsupervised learning, segmentation, ...)
- One of the major difficulties is in evaluating the success of a method
- validity depends on goals
  - if goal is to find 'interesting' clusters, this is rather difficult to quantify
  - external assessment: compare clustering to *a priori* clustering
  - relative assessment: compare one clustering methods results to another methods
  - for probabilistic methods, probabilistic measures for validating models

### ● ● ● | Choosing an Algorithm

- As we will see, different algorithms will result in clusters of different 'shapes'
- The appropriate shape will depend on the application and should be consider when choosing an algorithm
- match method to objectives

### ● ● ● | Families of Clustering Algorithms

- Partition-based methods
  - e.g., K-means
- Hierarchical clustering
  - e.g., hierarchical agglomerative clustering
- Probabilistic model-based clustering
  - e.g., mixture models
- Graph-based Methods
  - e.g., spectral methods

## Partition-based Clustering Algorithms

- Given set of  $n$  data points  $D = \{x(1), \dots, x(n)\}$  partition data into  $k$  clusters  $C = \{C_1, \dots, C_k\}$  such that each  $x(i)$  is assigned to a unique  $C_j$  and  $\text{score}(C, D)$  is minimized/maximized
- Combinatorial optimization: searching for allocation of  $n$  objects into  $k$  classes that maximizes score function
  - Number of possible allocations  $\approx n^k$
  - exhaustive typically finding the optimal solution is intractable
  - Resort to iterative improvement

## Generic Score Function

- Score function:
  - clusters compact  $\Rightarrow$  minimize within cluster distance,  $wc(C)$
  - clusters should be far apart  $\Rightarrow$  maximize distance between clusters,  $bc(C)$
- Given a clustering  $C$ , assign cluster centers,  $c_k$ 
  - if points belong to space means make sense, we can use the centroid of the points in the cluster:
 
$$c_k = \frac{1}{n_k} \sum_{x \in C_k} x$$
- $wc(C)$  = sum-of-squares within cluster distance
 
$$wc(C) = \sum_{k=1}^k wc(C_k) = \sum_{k=1}^k \sum_{x \in C_k} d(x, c_k)$$
- $bc(C)$  = distance between clusters
 
$$bc(C) = \sum_{1 \leq j < k \leq K} d(c_j, c_k)$$
- Score  $(C, D) = f(wc(C), bc(C))$

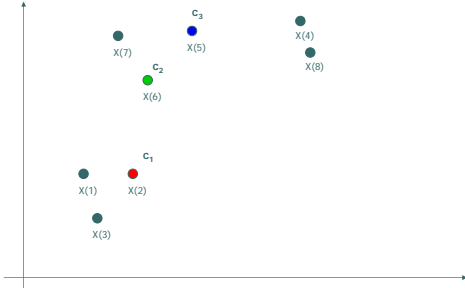
## CA#1: K-means

- Idea:
  - Start with randomly chosen cluster centers
  - Assign points to give greatest increase in score
  - Recompute cluster centers
  - Reassign points
  - Repeat until no changes

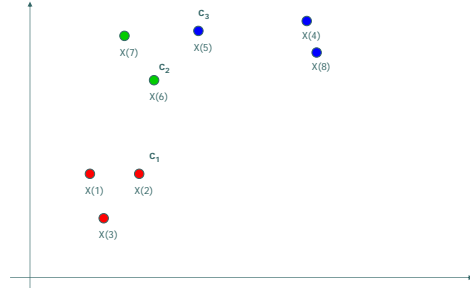
## K-means example

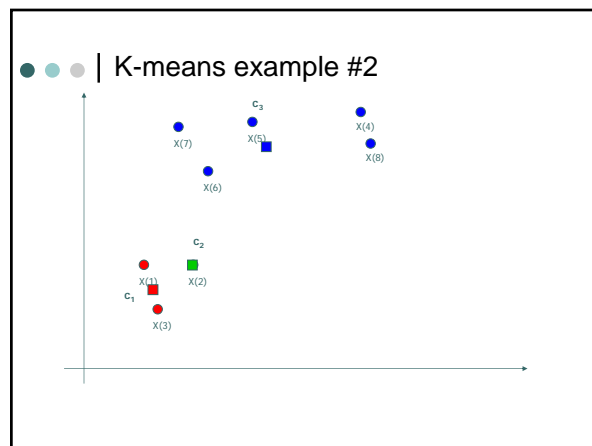
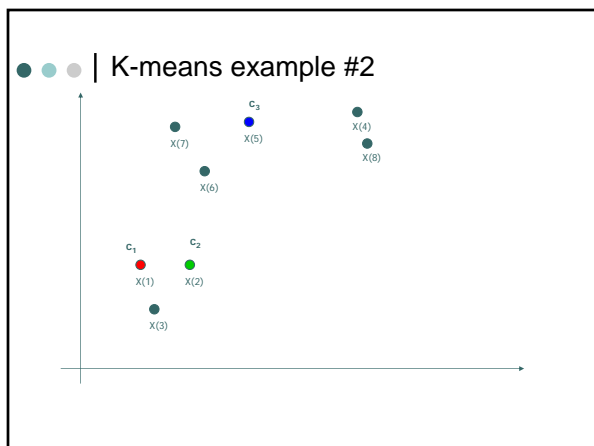
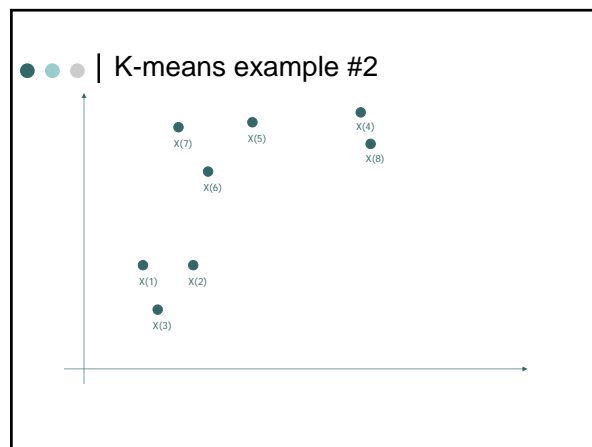
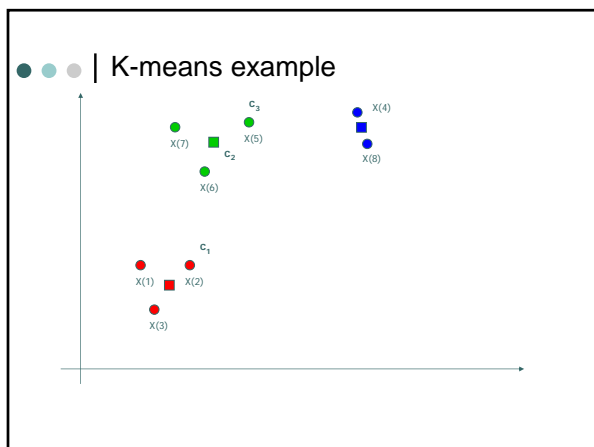
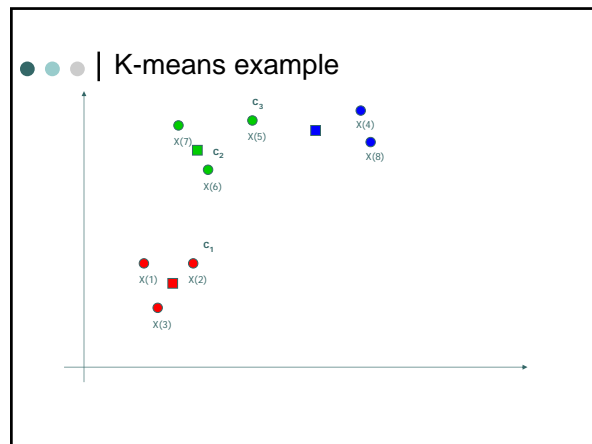
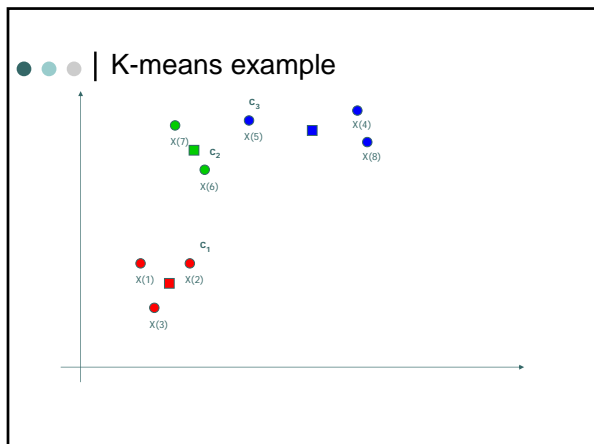


## K-means example



## K-means example





## ● ● ● | Demos

[k-means applet](#)

[image example](#)

## ● ● ● | Complexity

- Does algorithm terminate?
- Does algorithm converge to optimal solution?
- Time complexity one iteration?  $nk$

## ● ● ● | Algorithm Variations

- recompute centroid as soon as a point is reassigned
- allow merge and split of clusters
- methods for improving solution accuracy?
- in cases where means do not make sense
  - k-medoids – use one of the data points as center
  - categorical data -
- what if data set is too large for algorithm to be tractable?
  - compress data by replacing groups of objects by 'condensed representation'

## ● ● ● | An Incremental Clustering Algorithm

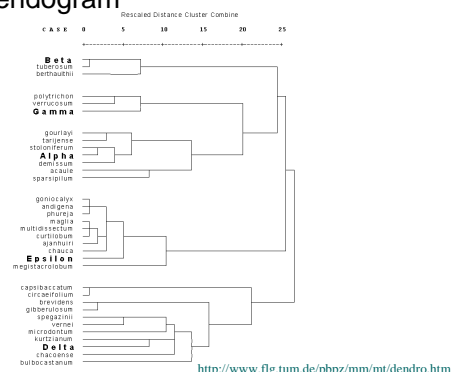
1. Assign first data point to a cluster
2. Consider next data point. Either assign data point to an existing cluster or create a new cluster. Assignment to cluster based on threshold
3. Repeat step 2 until all points are clustered

Useful for efficient clustering

## ● ● ● | CA #2: Hierarchical Clustering

- rather than deciding the number of clusters  $K$  at the start, build a hierarchy of nested clusters
- either gradually
  - merge points (agglomerative)
  - divide superclusters (divisive)
- result of either approach can be shown as a dendrogram which depicts the sequence of merges or splits

## ● ● ● | Dendrogram



## ● ● ● | Agglomerative Methods

- based on measures of distance between clusters

```
for i = 1 to n
  let  $C_i = \{x(i)\}$ 
while there is more than one cluster left do
  let  $C_i$  and  $C_j$  be the pair of clusters with minimum
   $D(C_i, C_j)$ 
   $C_i = C_i \cup C_j$ 
  remove  $C_j$ 
end
```

- time complexity?
- space complexity?

## ● ● ● | Measuring Distances between Clusters

- **single link**/nearest neighbor method:  
 $D(C_i, C_j) = \min\{d(x, y) \mid x \in C_i, y \in C_j\}$
- **complete link**/furthest neighbor method:  
 $D(C_i, C_j) = \max\{d(x, y) \mid x \in C_i, y \in C_j\}$
- **average link**:  
 $D(C_i, C_j) = \text{avg}\{d(x, y) \mid x \in C_i, y \in C_j\}$
- **centroid measure**:  
 $D(C_i, C_j) = d(c_i, c_j)$  where  $c_i$  and  $c_j$  are centroids
- **Ward's measure**: difference between total within cluster sum of squares for the two clusters separately and the sum of squares error in the merged cluster

## ● ● ● | Divisive Methods

- Begin with a single cluster, consisting of all the data points
- split into components
- ultimately ends with a partition in which each cluster has a single point
- **monolithic** methods split cluster using one variable at a time
- **polythetic** methods make splits based on all of the variables together; difficulty comes in how to choose potential splits
- in general, divisive methods are less widely used than agglomerative methods

## ● ● ● | Demos

- [http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial\\_html/AppletH.html](http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletH.html)
- <http://www.biology.ualberta.ca/jbrzusto/cluster.php#ClusterCalc>

## ● ● ● | Next Time

- Probabilistic Model-based Clustering
- Spectral Clustering