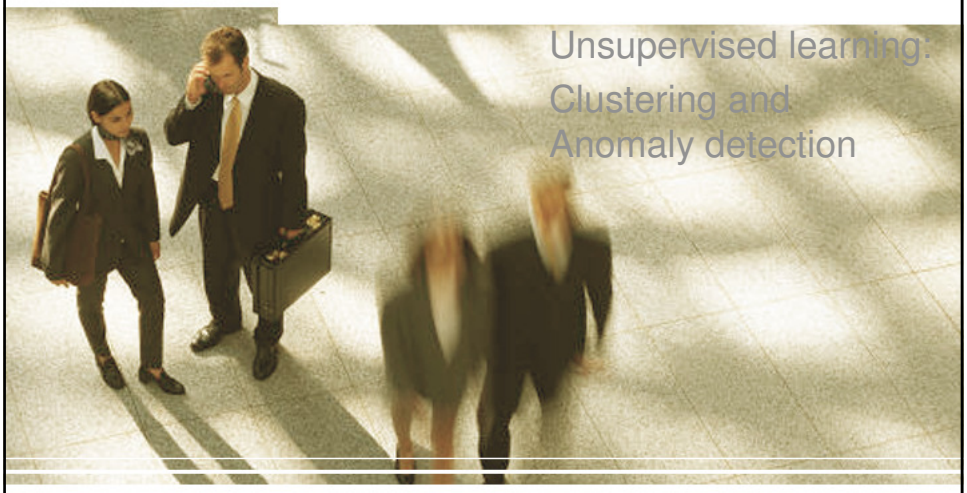
Class 12



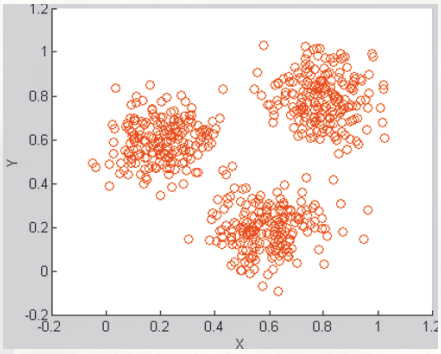
Unsupervised learning:
Clustering and
Anomaly detection

Learning Objectives

- Clustering
- Anomaly detection
- Algorithms
- Practical considerations

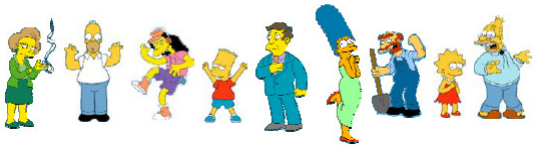
Unsupervised Learning: Clustering

- Given:
  - Data Set D (training set)
  - Similarity/distance metric/information
- Find:
  - Partitioning of data
  - Groups of similar/close items



Not a well-defined problem

What is a natural grouping among these objects?




Simpson's Family


School Employees


Females


Males

Similarity?

- Groups of similar customers
  - Similar demographics
  - Similar buying behavior
  - Similar health
- Similar products
  - Similar cost
  - Similar function
  - Similar store
  - …
- Similarity usually is domain/problem specific

Distance Between Records

- *d*-dim vector space representation and distance metric

r_1: 57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0

r_2: 78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0

…

r_N: 18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

Distance (r_1,r_2) = ???

- Pairwise distances between points (no *d*-dim space)
  - Similarity/dissimilarity matrix (upper or lower diagonal)
  - Distance: 0 = near, ∞ = far
  - Similarity: 0 = far, ∞ = near

```
-- 1 2 3 4 5 6 7 8 9 10
1  - d d d d d d d d d
2    - d d d d d d d d
3      - d d d d d d d
4        - d d d d d d
5          - d d d d d
6            - d d d d
7              - d d d
8                - d d
9                  - d
```

Properties of Distances: Metric Spaces

- A metric space is a set *S* with a global distance function *d*. For every two points *x, y* in *S*, the distance *d(x,y)* is a nonnegative real number.
- A metric space must also satisfy
  - d(x,y) = 0 iff *x = y*
  - d(x,y) = d(y,x) (symmetry)
  - d(x,y) + d(y,z) >= d(x,z) (triangle inequality)

Minkowski Distance (L_p Norm)

- Consider two records $x=(x_1,\ldots,x_d)$, $y=(y_1,\ldots,y_d)$:

$$d(x,y)=\sqrt[p]{|x_1-y_1|^p+|x_2-y_2|^p+\ldots+|x_d-y_d|^p}$$

Special cases:

- p=1: Manhattan distance

$$d(x,y)=|x_1-y_1|+|x_2-y_2|+\ldots+|x_p-y_p|$$

- p=2: Euclidean distance

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_d-y_d)^2}$$

Only Binary Variables

2x2 Table:

	0	1	Sum
0	a	b	a+b
1	c	d	c+d
Sum	a+c	b+d	a+b+c+d

- Simple matching coefficient: (symmetric) $d(x, y) = \frac{b + c}{a + b + c + d}$
- Jaccard coefficient: (asymmetric) $d(x, y) = \frac{b + c}{b + c + d}$

Mixtures of Variables

- Weigh each variable differently
- Can take “importance” of variable into account (although usually hard to quantify in practice)

Clustering: Informal Problem Definition

Input:

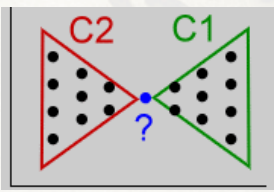
- A data set of *N* records each given as a *d*-dimensional data feature vector.

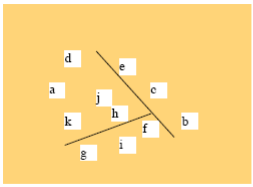
Output:

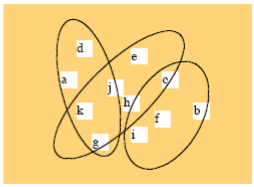
- Determine a natural, useful "partitioning" of the data set into a number of (k) clusters and noise such that we have:
  - High similarity of records within each cluster (intra-cluster similarity)
  - Low similarity of records between clusters (inter-cluster similarity)

Types of Clustering

- Hard Clustering:
  - Each object is in one and only one cluster
- Soft Clustering:
  - Each object has a probability of being in each cluster







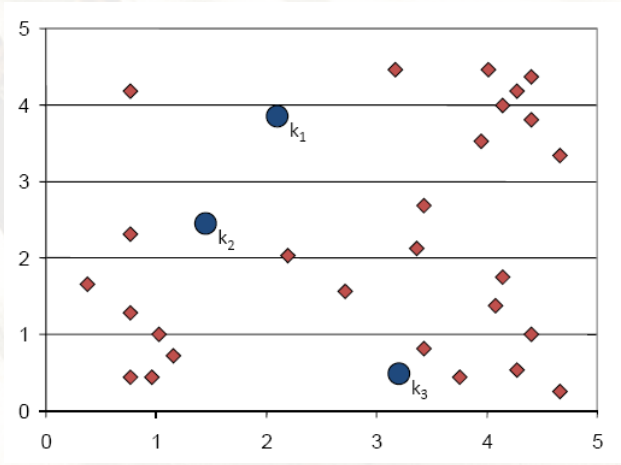
Clustering Algorithms

- Partitioning-based clustering
  - K-means clustering
  - K-medoids clustering
  - EM (expectation maximization) clustering
- Hierarchical clustering
  - Divisive clustering (top down)
  - Agglomerative clustering (bottom up)
- Density-Based Methods
  - Regions of dense points separated by sparser regions of relatively low density

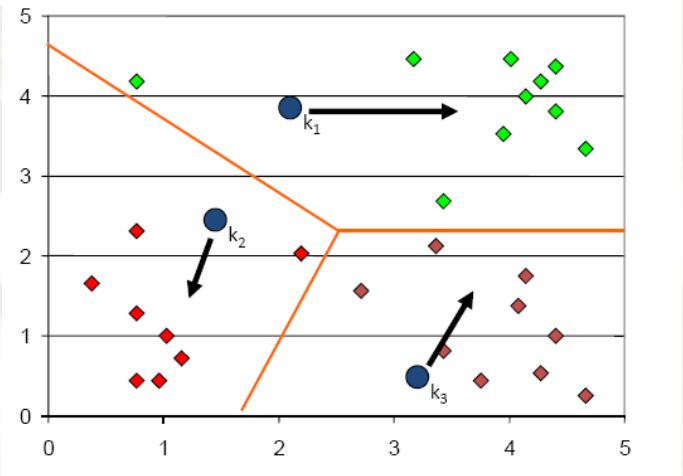
K-Means

1. Decide on a value for *k.*
2. Initialize the *k cluster centers (randomly, if* necessary).
3. Decide the class memberships of the *N objects by* assigning them to the nearest cluster center.
4. Re-estimate the *k cluster centers, by assuming the* memberships found above are correct.
5. If none of the *N objects changed membership in* the last iteration, exit. Otherwise goto 3.

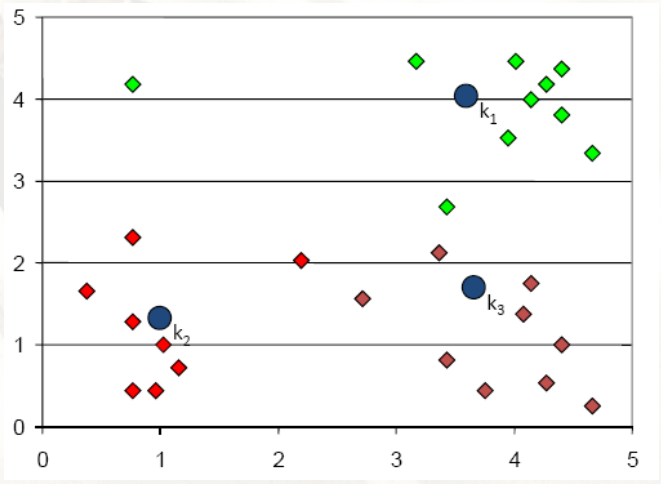
K-Means: Step 1



K-Means: Step 2

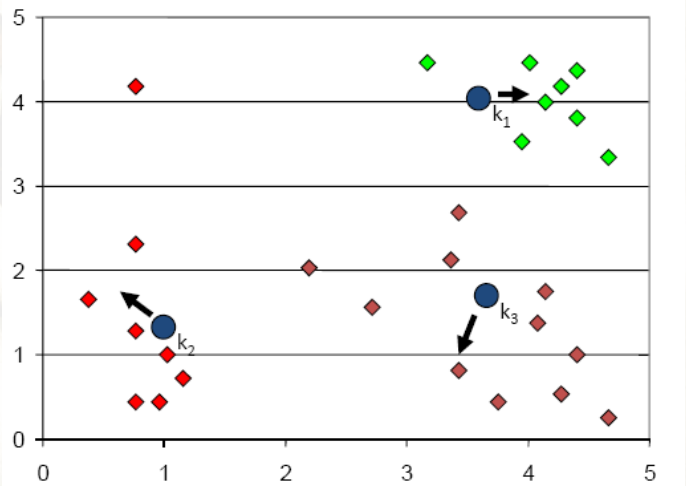


K-Means: Step 3



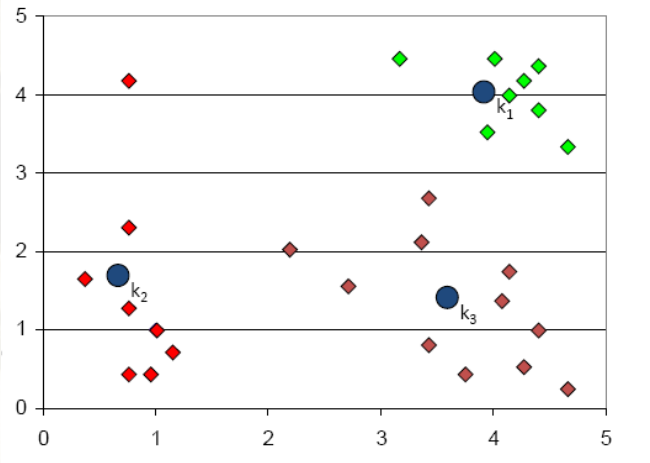
17

K-Means: Step 4

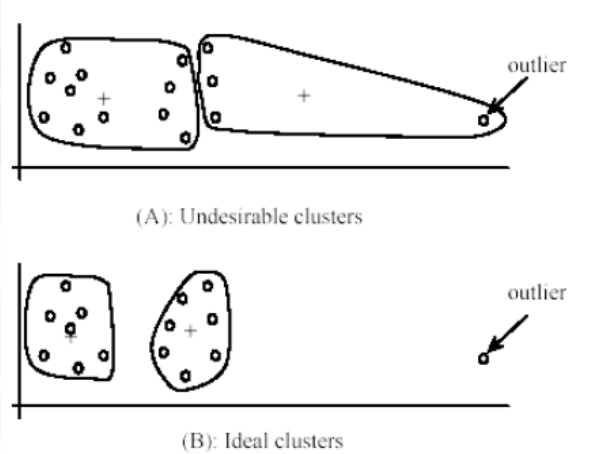


18

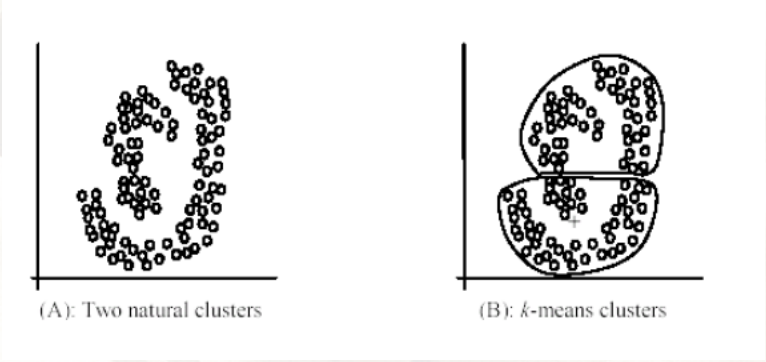
K-Means: Step 5



K-Means is sensitive to outliers



K-Means and complex clusters



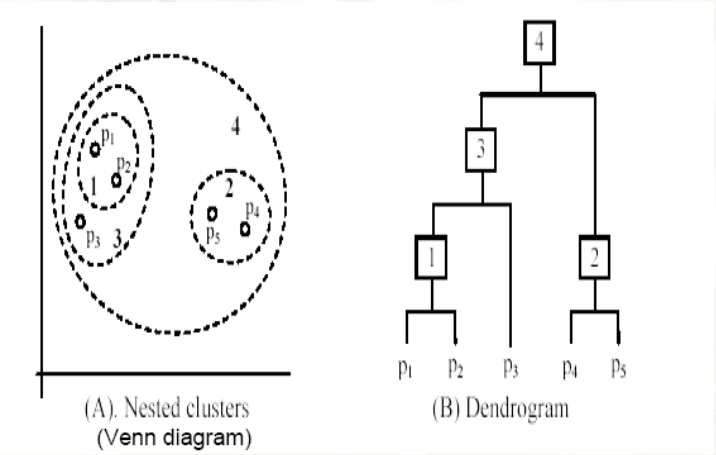
K-Means: Summary

- Despite its weaknesses, *k-means is still the most popular* algorithm due to its simplicity and efficiency
- Other clustering algorithms have also their own weaknesses
  - No clear evidence that any other clustering algorithm performs better than *k-means in general*
- Some clustering algorithms may be more suitable for some specific types of dataset, or for some specific application problems, than the others
  - Comparing the performance of different clustering algorithms is a difficult task
- No one knows the correct clusters!

Hierarchical clustering

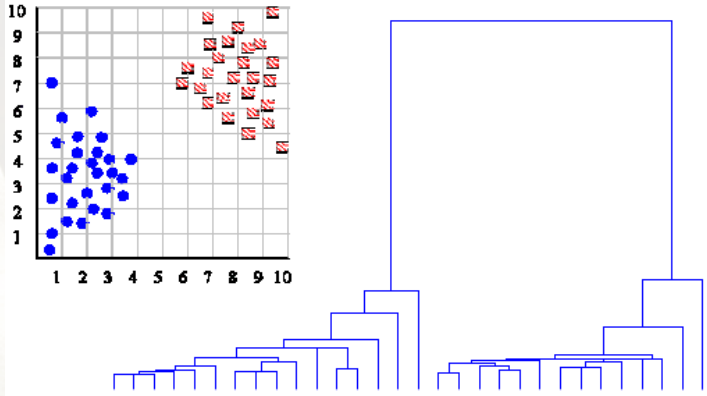
- Hierarchical agglomerative (bottom-up) clustering builds the dendrogram from the bottom level
- The algorithm
  - At the beginning, each instance forms a cluster (also called a node)
  - Merge *the most similar (nearest) pair of clusters*
    - i.e., The pair of clusters that have *the least distance among all* the possible pairs
  - Continue the merging process
  - Stop when all the instances are merged into a single cluster (i.e., the *root cluster)*

Example



Determining the number of clusters

2 highly separated subtrees => 2 clusters



Hierarchical clustering: summary

- No need to specify the number of clusters in advance.
- Hierarchal nature maps nicely onto human intuition for some domains
- They do not scale well: time complexity of at least O(n^2), *where n is the number of total objects.*
- Like any heuristic search algorithms, local optima are a problem.
- Interpretation of results is (very) subjective

Clustering: practical issues

- What is a cluster?
- Which features and normalization scheme?
- How to define pair-wise similarity?
- How many clusters?
- Which clustering method?
- Does the data have any clustering tendency?
- Are the discovered clusters & partition valid?

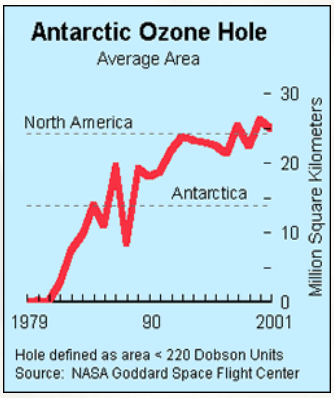
Anomaly (outlier) detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Variants of Anomaly/Outlier Detection Problems
  - Given a database D, find all the data points $\mathbf{x} \in$ D with anomaly scores greater than some threshold t
  - Given a database D, find all the data points $\mathbf{x} \in$ D having the top-n largest anomaly scores f($\mathbf{x}$)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a test point $\mathbf{x}$, compute the anomaly score of $\mathbf{x}$ with respect to D
- Applications:
  - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection

Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels
- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?
- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



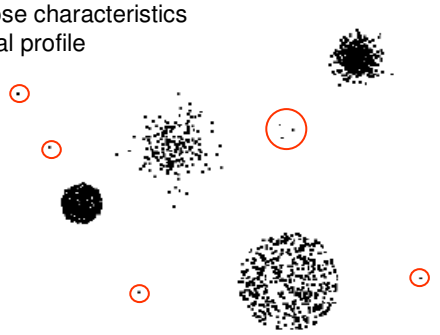
Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Anomaly Detection

- Challenges
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging (just like for clustering)
  - Finding needle in a haystack
- Working assumption:
  - There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data

Anomaly Detection Schemes

- General Steps
  - Build a profile of the "normal" behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the "normal" profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile

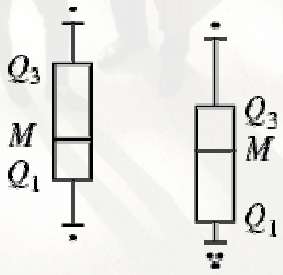


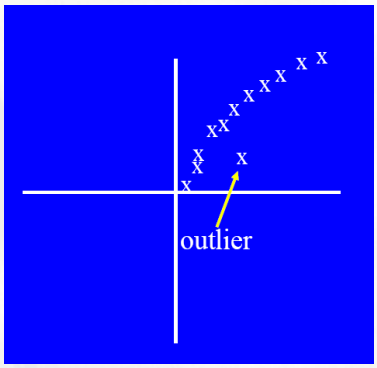
- Types of anomaly detection schemes
  - Graphical & Statistical-based
  - Distance-based
  - Model-based

Graphical Approaches

- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)

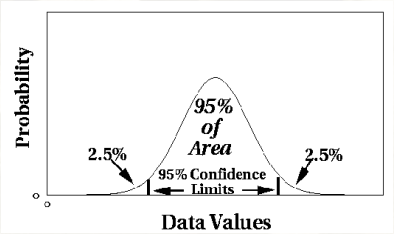
- Limitations
  - Time consuming
  - Subjective





Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

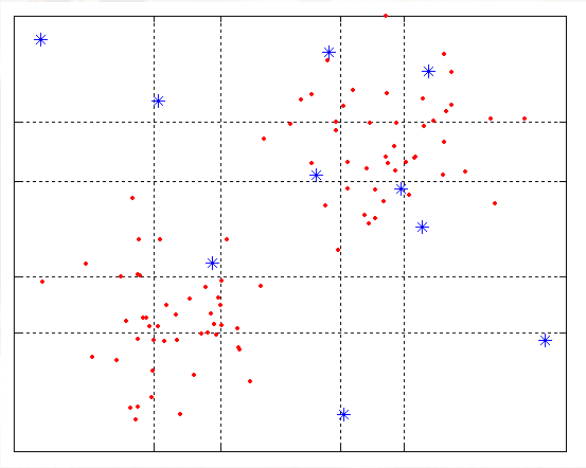
Distance-based Approaches

- Data is represented as a vector of features
- Two major approaches
  - Nearest-neighbor based
  - Clustering based

Nearest-Neighbor Based Approach

- Approach:
  - Compute the distance between every pair of data points
  - There are various ways to define outliers:
    - Data points for which there are fewer than *p* neighboring points within a distance *D*
    - The top n data points whose distance to the kth nearest neighbor is greatest
    - The top n data points whose average distance to the k nearest neighbors is greatest

Example



Clustering-Based

- Basic idea:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, they are outliers