Data Warehousing & Data Mining

- **Clustering – I**

Lecture 11 Dated 20/12/2010

In this lecture

- The Problem of Clustering
- Types of Clustering
- Similarity and Dissimilarity
- Distance Measures
- Scales of Measurement
- Various Distance Functions

- The lecture is based (and adapted) from
  – "CS345 --- Lecture Notes", by Jeff D Ullman at Stanford. http://www-db.stanford.edu/~ullman/cs345-notes.html
  – Vipin Kumar's course in data mining offered at University of Minnesota
  – official text book slides of Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, August 2000

2

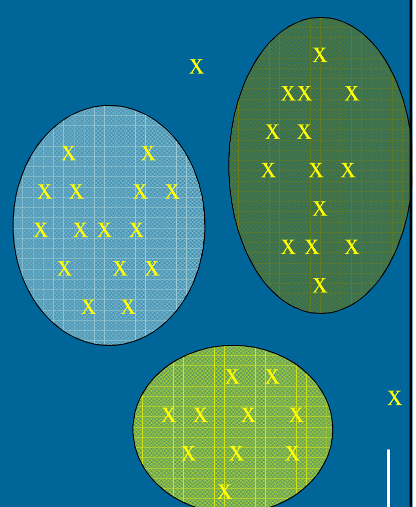
The Problem of Clustering

- Given a set of points, with a notion of distance between points, group the points into some number of *clusters*, so that members of a cluster are in some sense as nearby as possible.
- Clustering is unsupervised classification: no predefined classes.
- Formally, Clustering is the process of grouping data points such as intra-cluster distance is minimized and inter-cluster distance is maximized.

3

Example Applications

- Marketing: Help marketers discover distinct groups in their customer bases
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location



4

What is not Cluster Analysis?

- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
- Graph partitioning
  - Some mutual relevance and synergy, but areas are not identical

5

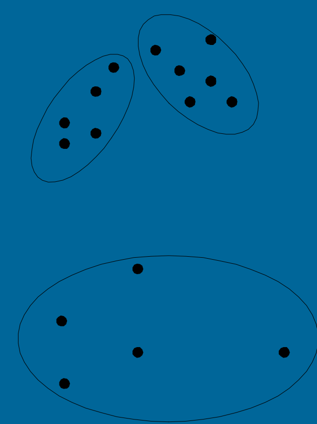
Types of Clustering

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
  - Partitional Clustering
    - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
  - Hierarchical clustering
    - A set of nested clusters organized as a hierarchical tree
- Other distinctions – *coming slides*

6

Partitional Clustering

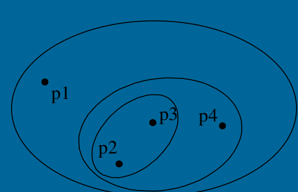
Original Points



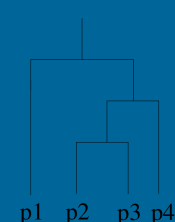
A Partitional Clustering

7

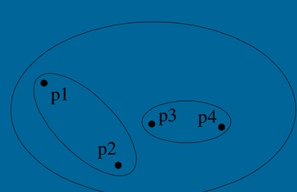
Hierarchical Clustering



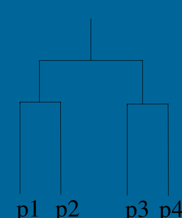
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

8

Other Distinctions Between Sets of Clusters

- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or 'border' points
- Fuzzy versus non-fuzzy
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- Partial versus complete
  - In some cases, we only want to cluster some of the data
- Heterogeneous versus homogeneous
  - Cluster of widely different sizes, shapes, and densities

9

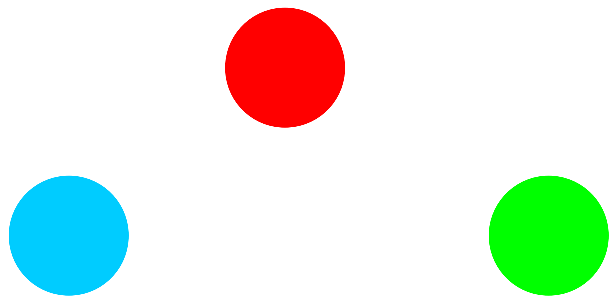
Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

10

Types of Clusters: Well-Separated

- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

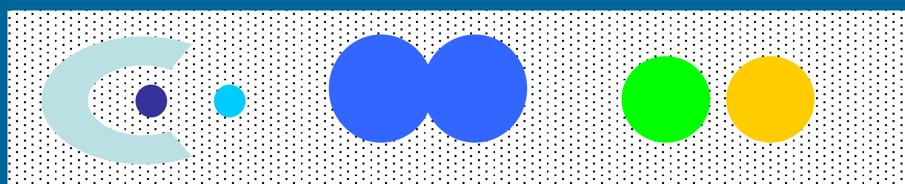
Types of Clusters: Center-Based

- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster

4 center-based clusters

Types of Clusters: Density-Based

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Types of Clusters: Objective Function …

- Map the clustering problem to a different domain and solve a related problem in that domain
  - Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points
  - Clustering is equivalent to breaking the graph into connected components, one for each cluster.
  - Want to minimize the edge weight between clusters and maximize the edge weight within clusters

Characteristics of the Input Data Are Important

- Type of proximity or density measure
  - This is a derived measure, but central to clustering
- Sparseness
  - Dictates type of similarity
  - Adds to efficiency
- Type of Data
  - Dictates type of similarity
  - Other characteristics, e.g., autocorrelation
- Dimensionality
- Noise and Outliers
- Type of Distribution

15

Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

16

Distance Measures

- Each clustering problem is based on some kind of "distance" between points.
  - Distance between documents
  - Distance between demographic details of two customers
  - Distance between transactions
  - Distance between strings (proteins, addresses etc.)
- Two major classes of distance measure:
  1. *Euclidean* : based on position of points in some *k* -dimensional space.
  2. *Noneuclidean* : not related to position or space.

17

Scales of Measurement

- Applying a distance measure largely depends on the type of input data
- Major scales of measurement:
  1. **Nominal Data (aka Nominal Scale Variables)**
     - Typically classification data, e.g. m/f
     - no ordering, e.g. it makes no sense to state that M > F
     - Binary variables are a special case of Nominal scale variables.
  2. **Ordinal Data (aka Ordinal Scale)**
     - ordered but differences between values are not important
     - e.g., political parties on left to right spectrum given labels 0, 1, 2
     - e.g., Likert scales, rank on a scale of 1..5 your degree of satisfaction
     - e.g., restaurant ratings

18

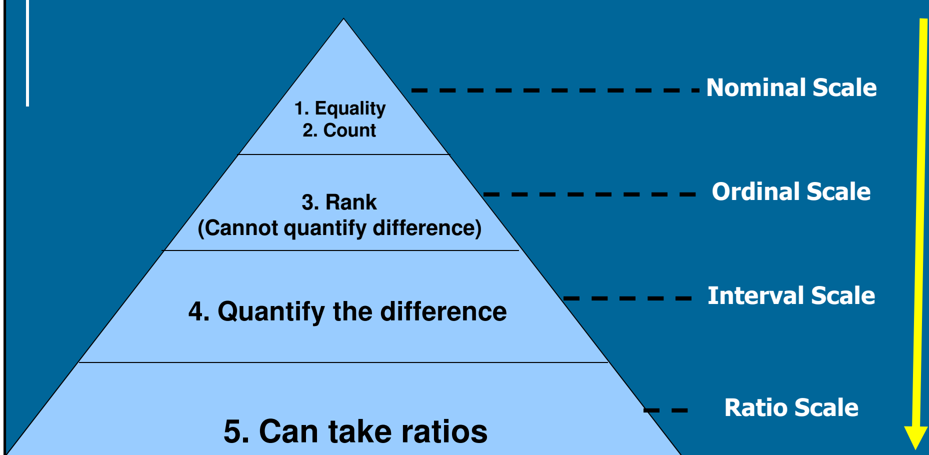
Scales of Measurement

- Applying a distance function largely depends on the type of input data
- Major scales of measurement:

3. **Interval Data (aka interval scaled)**
    - Ordered and equal intervals. Measured on a linear scale.
    - Differences make sense
    - e.g., temperature (C,F), dates
4. **Ratio Data (aka ratio scaled)**
    - Continuous positive measurements on a nonlinear scale
    - Ordered
    - e.g., height, weight, age, length

Scales of Measurement

- Only certain operations can be performed on certain scales of measurement.



Axioms of a Distance Measure

- *d* is a distance measure if it is a function from pairs of points to reals such that:
  1. d(x,x) = 0.
  2. d(x,y) = d(y,x).
  3. d(x,y) $\geq$ 0.
  4. d(x,y) $\leq$ d(x,z) + d(z,y) (triangle inequality).

21

Some Euclidean Distances

- *L_2 norm* (also common or Euclidean distance):

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

  – The most common notion of "distance."

- *L_1 norm* (also Manhattan distance)
  – distance if you had to travel along coordinates only.

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

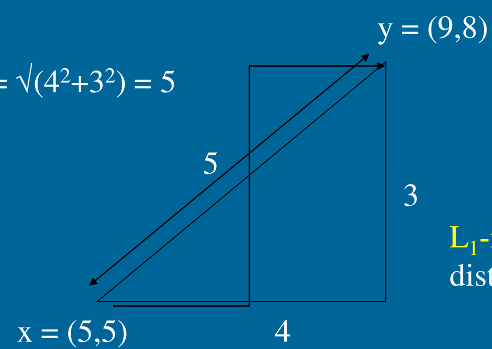
- Both norms are special forms of Minwoski norm

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

22

Examples L_1 and L_2 norms

L_2-norm:
dist(x,y) = $\sqrt{(4^2+3^2)} = 5$



L_1-norm:
dist(x,y) = 4+3 = 7

23

Another Euclidean Distance

- *L_∞ norm* : d(x,y) = the maximum of the differences between *x* and *y* in any dimension.
- Note: the maximum is the limit as *n* goes to ∞ of what you get by taking the *n* th power of the differences, summing and taking the *n* th root.

24

Non-Euclidean Distances

- *Jaccard measure* for binary vectors
- *Cosine measure* = angle between vectors from the origin to the points in question.
- *Edit distance* = number of inserts and deletes to change one string into another.

25

Jaccard Measure

- A note about Binary variables first
  - **Symmetric binary variable**
    - If both states are equally valuable and carry the same weight, that is, there is no preference on which outcome should be coded as 0 or 1.
    - Like "gender" having the states male and female
  - Asymmetric **binary variable**:
    - If the outcomes of the states are not equally important, such as the positive and negative outcomes of a disease test.
    - We should code the rarest one by 1 (e.g., HIV positive), and the other by 0 (HIV negative).
  - Given two asymmetric **binary** variables, the agreement of two 1s (a positive match) is then considered more important than that of two 0s (a negative match).

26

Edit Distance

- The edit distance of two strings is the number of inserts and deletes of characters needed to turn one into the other.
- Equivalently, d(x,y) = |x| + |y| -2|LCS(x,y)|.
  - LCS = *longest common subsequence* = longest string obtained both by deleting from *x* and deleting from *y*.

27

The Curse of Dimensionality

- While clustering looks intuitive in 2 dimensions, many applications involve 10 or 10,000 dimensions.
- High-dimensional spaces look different: the probability of random points being close drops quickly as the dimensionality grows.
- In a high dimension space, almost all pairs of points are about as far away as average.

28