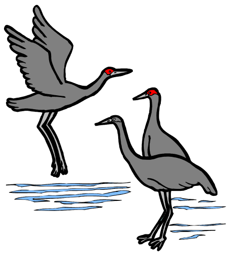


Clustering

Birds of a feather flock together.

Vikram Pudi
vikram@iiit.ac.in
IIIT Hyderabad

The Clustering Problem

Outlook	Temp (°F)	Humidity (%)	Windy?
sunny	75	70	true
sunny	80	90	true
sunny	85	85	false
sunny	72	95	false
sunny	69	70	false
overcast	72	90	true
overcast	73	88	true
overcast	64	65	true
overcast	81	75	false
rain	71	80	true
rain	65	70	true
rain	75	80	false
rain	68	80	false
rain	70	96	false

Find groups of similar records.

Need a function to compute similarity, given 2 input records

⇒ Unsupervised learning

Applications

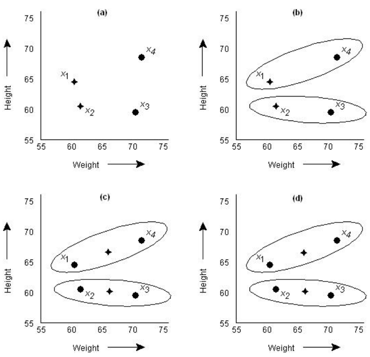
- Targetting similar people or objects
  - Student tutorial groups
  - Hobby groups
  - Health support groups
  - Customer groups for marketing
  - Organizing e-mail
- Spatial clustering
  - Exam centres
  - Locations for a business chain
  - Planning a political strategy

Measurement of similarity

- Nominal (categorical) variables
  d(x,y) = 1 - m/n
  m = no of matches among n attributes, or
  m = sum of weights of matching attributes, and n is the sum of weights of all attributes
- Numeric variables
  - Euclidean, manhattan, minkowski,…
  - Ordinal
    - z = (rank-1)/(M-1) where M is maximum rank
- Above are examples
  - Similarity is ultimately application dependent
  - Requires various kinds of preprocessing
    - Scaling: Convert all attributes to have same range
    - z-score: z = (value-mean)/m
      where m is the mean absolute deviation

Partitioning technique: *k*-Means

- Initial *k* means = random records
- Iterate as long as clusters change:
  - Put each record X in the cluster to whose mean it is closest
  - Recompute means as the average of all points in each cluster



Evaluating Clustering Quality

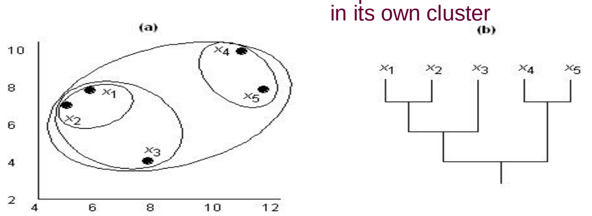
- Minimize squared error
  Here m_i is the mean (or other centre) of cluster *i*
- Can also use absolute error
- Can be used to find best initial random means in *k*-means.

$$E = \sum_{i=1}^{N} \sum_{x \in C_i} d(x, m_i)^2$$

Hierarchical Methods

Agglomerative (e.g. AGNES):
- Start: Each point in separate cluster
- Merge 2 closest clusters
- Repeat until all records are in 1 cluster.

Divisive (e.g. DIANA)
- Start: All points in 1 cluster
- Find most extreme points in each cluster.
- Regroup points based on closest extreme point
- Repeat until each record is in its own cluster

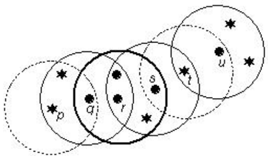


Measuring Cluster Distances

- Single link: Minimum distance
- Complete link: Maximum distance
- Average link: Average distance
- *Mean link*: Mean distance

Density-based Methods: e.g. DBSCAN

- Neighbourhood: Records within distance of ε from given record.
- Core point: Record whose neighbourhood contains at least μ records.
- Find all core points and create a cluster for each of them.
- If core point Y is in the neighbourhood of core point X, then merge the clusters of X and Y.
- Repeat above step for all core points until clusters do not change.



Mining Outliers using Clustering

- Outliers are data points that deviate significantly from the norm.
- Useful in fraud detection, error detection (in data cleaning), etc.
- Technique:
  - Apply any clustering algorithm
  - Treat clusters of very small size as containing only outliers