Cluster Analysis

7.1 – 7.3

What is Cluster Analysis?

- Cluster – collection of data objects similar to others in the cluster and dissimilar to objects in other clusters
- No training data
  - costly to collect and label
  - labels are unknown
  - AI – learning by observation; unsupervised learning

Clustering Applications

- Marketing
  - discovery of customer groups/demographics
- Biology
  - taxonomy derivation, gene categorization
- Outlier detection
  - Credit card fraud, network intrusion
- Data Preprocessing
  - Training data preparation

Clustering Challenges

- Scalability
  - work with entire data set
- Different data types
  - interval, binary, categorical, ordinal
- Clusters of arbitrary shape
- Input parameters for clustering algorithms
  - number of clusters, size
  - output highly sensitive
  - burden on users

Clustering Challenges

- Noisy data
- Insensitivity to incremental record entry
  - Incorporating into existing clusters vs clustering from scratch
- Insensitivity to record order
- High dimensionality
- Constraint inputs
- Usable, interpretable results

Data Structures

- Data matrix
  - *n* objects, *p* variables – two mode
  - Representation: (n x p) matrix
- Dissimilarity matrix
  - used by most clustering algorithms
  - *n* objects – one mode
  - Representation: (n x n) lower triangular matrix
    - cell (i, j) is the dissimilarity between object i and object j

Interval Scaled Variables

- Continuous numerical measures on a linear scale
- Standardization
  - Units matter – smaller scales can give larger weight
  - Weighting may be determined by user
  - z-score with mean absolute deviation
    - mean absolute deviation - $s = (1/n) * \Sigma (|x_i - mean_i|)$
    - $z\text{-}score_i = x_i - mean_i / s$
    - Standard deviation squares the difference – exaggerates outliers

Interval Scaled Variables

- Euclidian distance
  - $d(i, j)=\sqrt{((x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn})^2)}$
- Manhattan (city block) distance
  - $d(i, j)=|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}|$
- Minkowski distance
  - $d(i, j)=(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{in} - x_{jn}|^p)^{1/p}$

Binary Variables

- Two states – simplify to 0, 1
- Dissimilarity matrix
  - q = 1 in both; r = 1 in i, 0 in j;
    s = 0 in i, 1 in j; t = 0 in both
- Symmetric binary – both states of equal interest
  - d(i, j) = (r+s)/(q+r+s+t)
- Asymmetric binary – one state more interesting
  - d(i, j) = (r+s)/(q+r+s)
  - similarity(i, j) = 1 – d(i, j) = q/(q+r+s)

Categorical Variables

- Multiple distinct states
- d(i, j) = (p-m)/p
  - m = number of variables in i and j that match
  - p = total number of variables

Ordinal Variables

- Categorical where states are in a meaningful sequence – ranking
- Map each value to its rank, r = (1, 2, …, M)
- Scale to (0.0, 1.0), $z = (r - 1)/(M - 1)$
- Compute d(i, j) with interval-scaled methods

Ratio-Scaled Variables

- Nonlinear scale
  - Population Growth, Radioactive Decay = Ae^{Bt}
- Logarithmic Transformation
  - $y = \log_a x$
  - Treat y as interval valued
- Ordinal Data
  - Rank and interval-scale

Mixed and Vector Variables

- Mixed types – normalize all to (0.0, 1.0) and process all variables together
- Vectors
  - cosine similarity measure
    - $s(x, y) = (x^t \cdot y)/(||x||\ ||y||)$
  - Tanimoto coefficient
    - $s(x, y) = (x^t \cdot y)/(x^t \cdot x + y^t \cdot y - x^t \cdot y)$

Partitioning Methods

- k partitions of n objects; $k \leq m$
- Each group has at least one object; each object in only one group
- Iterative relocation
  - Initial partitioning; relocate objects between groups
  - "good" partitioning – objects in same cluster are near, objects of different clusters are far apart
  - k-means – cluster represented by mean of the objects
  - k-medoids – cluster represented by object near center
- "Spherical" clusters, small or medium-sized databases

Hierarchical Methods

- Agglomerative – bottom-up
  - Start with each object as its own cluster; merge clusters near each other; stop when one cluster exists
- Divisive – top-down
  - Start with all objects in one cluster; split into clusters furthest apart; stop when each object is its own cluster
- Splits/merges cannot be undone if done wrong

Density-Based Methods

- Increase size of a given cluster until density is below a set level
- For each data point in cluster, a minimum number of other points must exist in a certain radius
- Better at finding non-circular clusters

Grid and Model Based Methods

- Grid-Based
  - Quantize object space into finite number of cells
  - Cluster by including or excluding cells
  - Fast computation – dependent on number of cells
- Model-Based
  - Hypothesize a model of each cluster
    - Density function, statistics analysis
  - Find best fit of data to model

High-Dimensional Methods

- Relevancy of dimensions
- High dimensions results in sparse data
  - Larger overall distances
  - Low density
- Subspace Clustering
  - Cluster in significant subsets of dimensions
- Frequent Pattern Clustering
  - Cluster on frequent patterns among subsets of dimensions

Constraint-Based Methods

- Constraint
  - user expectation or application-specific property of resulting clusters
- Use constraints specified by user or application
  - Starting point for clustering
  - Refine quality of resulting clusters