Data Clustering

Michael J. Watts

http://mike.watts.net.nz

Lecture Outline

- Data clustering
- Distance measures
- K-Means clustering
- Hierarchical clustering

Data Clustering

- Grouping similar items together
- Assigns items into subsets
  - Usually single subsets
  - Usually uses distance measures
- Vector quantisation
- Data reduction

Distance Measures

- Many distance measure exist
  - Euclidean
  - Manhattan
  - Mahalanobis
- Each useful in different ways

k-Means Clustering

- Simple clustering technique
- Widely used
- Requires *a priori* selection of number of clusters
- Iterative, can be slow
- Depends on selection of clusters / seeds

k-Means Clustering

1) select *k* seed examples as initial centres
2) calculate the distance from each centre to each example
3) assign each example to the nearest cluster
4) calculate new centres
5) repeat steps 2-4 until a stopping condition is reached

Hierarchical Clustering

- Also called agglomerative hierarchical clustering
- Constructs a tree structure
  - Dendogram
- Close examples are lower down the tree
- Each cluster is treated as a single entity
- All clusters are grouped into a super-cluster
- Widely used in bioinformatics

Hierarchical Clustering

1) Calculate distance between each pair of vectors
2) Merge closest two vectors / entities
   - New entity has average of parents
3) Re-calculate the distances
4) Repeat 1-3 until all vectors / entities are merged

Summary

- Clustering groups together similar items
- Similarity is measured by distance
- Many distance measures exist
- k-Means groups examples into individual clusters
- Hierarchical groups examples into ever-smaller groups