Lance-Williams Algorithm

- D = D(j,i) distance between j and i for Iterations:
  - Initial: D = a, D = b + pair of closest clusters
  - Add D = c, D = d
  - Add D = e

Agglomerative clustering

- Start with collection C of n singleton clusters
  - Each element contains one data point
  - c(0) - c(n)

- Need to define distance metric over clusters
  - Slow O(n^2) - create, traverse distance matrix

- Repeat until only one cluster left:
  - Find pair of clusters to closest
  - Merge clusters and remove C from the collection

Agglomerative clustering: example

Hierarchical Clustering

- Selecting K - how many clusters is the appropriate number of clusters?
  - no clustering algorithm needs to pick K or density threshold
  - top-down approach - starts with all elements in one cluster
  - bottom-up approach - starts with individual elements

Hierarchical K-means

- Top-down approach:
  - can cross boundaries imposed by top levels
  - fast, recursive calls operate on a slice of K

Summary

- Clustering discover underlying sub-populations
  - K-means
  - hierarchical clustering: hierarchical tree of clusters
  - need to pick k for k-means
  - hierarchical algorithm needs to be least minimum
  - Ward's method: for each cluster, minimize sum of squares
  - Average linkage:
    - distance between clusters is the average of distances to all points in each cluster
  - Complete linkage:
    - distance between clusters is the maximum distance between a point in one cluster and a point in another cluster
  - Centroids:
    - average of cluster points
  - Distance between clusters: Euclidean distance

IAML: Hierarchical Clustering

Victor Tavrow, Nigel Godard
School of Informatics
Semester 1