Hierarchical Clustering

36-350: Data Mining
25 September 2006
Last time...

- Unsupervised learning problems; finding clusters
- K means
  - divide into k clusters to minimize within-cluster variance * cluster size
- local search, local minima
Limits of k-Means

- Local search can get stuck
- Random starts help
- Sum-of-squares likes ball-shaped clusters
- How to pick k?
- No relations between clusters
Hierarchical Clustering

- Basic idea: cluster the clusters
- High-level clusters contain multiple low-level clusters
- Clusters are now related
- Don’t need to chose k
- Assumes a hierarchy makes sense...
Ward’s Method

1. Start with every point in its own cluster

2. For each pair of clusters, calculate “merging cost” = increase in sum of squares

3. Merge least-costly pair

4. Stop when merging cost takes a big jump
Ward’s method applied to the images from lecture 3: ocean, tigers, flowers

Jump in merging cost suggests 3 clusters - almost exactly right ones, too (but thinks flower5 is a tiger)
• Don’t have to chose $k$

• Sum of squares is worse, generally, than $k$-means (for equal $k$)

• more constrained search

• prefers to merge small clusters, all else equal
Minimizing the mean distance from the center tends to make spheres, which can be silly.

note how Ward’s is less balanced.
Single-link clustering

1. Start with every point in its own cluster

2. Calculate gaps between every pair of clusters = distance between 2 closest points in each cluster

3. Merge clusters with smallest gap
Examples where single-link doesn’t work so well

- k-Means
- Ward’s
- Single-link