Midterm: Friday Mar 4 1:10-3pm
UC273
2-sided sheet & calculator

Hierarchical clustering

Start with n clusters (n observations)
- sucessing group together observations/cluster that are close.

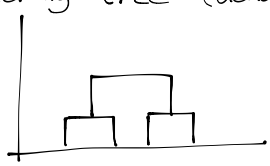
have 2 clusters U, V
$d(U,V) = ?$

Single linkage
$d(U,V) = \min(d(\underset{\sim}{x}_i, \underset{\sim}{x}_j) : \underset{\sim}{x}_i \in U, \underset{\sim}{x}_j \in V)$

Complete linkage
$d(U,V) = \max(d(\underset{\sim}{x}_i, \underset{\sim}{x}_j) : \underset{\sim}{x}_i \in U, \underset{\sim}{x}_j \in V)$

Average linkage
$d(U,V) = \text{ave}(d(\underset{\sim}{x}_i, \underset{\sim}{x}_j) : \underset{\sim}{x}_i \in U, \underset{\sim}{x}_j \in V)$

Result clustering tree (dendrogram)



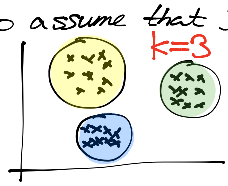
Blackboard
Clustering of atheletes record data

- dendrograms for 3 methods
- KOR clearly "different"

Other note:
- often do hierarchical clustering with "similarity" matrices
- high similarity = low distance
e.g. single linkage define similarity b/w clusters = max pairwise similarity

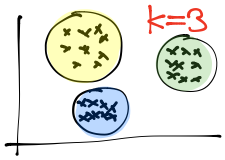
Towards model-based clustering

k-means clustering
Informal model: $\underset{\sim}{X}$ has density $f(\underset{\sim}{x}) = \underbrace{\theta_1 f_1(\underset{\sim}{x}) + \cdots + \theta_k f_k(\underset{\sim}{x})}_{\text{mixture model}}$
where $\theta_1, \ldots, \theta_k > 0$ with $\theta_1 + \cdots + \theta_k = 1$
- Also assume that $f_i(\underset{\sim}{x}) f_j(\underset{\sim}{x}) \doteq 0$ for $i \neq j$



- $f_1(\underset{\sim}{x}), \ldots, f_k(\underset{\sim}{x})$ represent density of sub-populations
- $\theta_1, \ldots, \theta_k$ represent sub-pop'n proportions.
- in practice, $f_1(\underset{\sim}{x}), \ldots, f_k(\underset{\sim}{x})$ and $\theta_1, \ldots, \theta_k$ are unknown
- If we know $f_1(\underset{\sim}{x}), \ldots, f_k(\underset{\sim}{x})$ then given an observation $\underset{\sim}{x}^*$, we can predict very well which sub-pop'n it belongs to.

Problem: Given data $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$, how to determine k clusters Corresponding to k sub-pop'ns) of observations.
- Start by assuming k is known.



– need to determine
- center points (centroids) of the k clusters
- determine shape of clusters
- which observations belong to which clusters

For arbitrary centroids $\mu_1, \cdots, \mu_k$, find groups of observations $\underbrace{G_1, G_2, \cdots G_k}_{\text{disjoint}}$ to minimize

$$\underbrace{\sum_{j=1}^{k} \sum_{i \in G_j}}_{\text{minimize}} d(\underset{\sim}{x}_i, \underset{\sim}{\mu}_j) \longrightarrow \text{fixed}$$

$\Rightarrow G_1^*, G_2^*, \cdots, G_k^*$ depend on $\mu_1, \cdots, \mu_k$

Now minimize (w.r.t. $\mu_1, \cdots, \mu_k$),

$$\sum_{j=1}^{k} \sum_{i \in G_j^*} d(\underset{\sim}{x}_i, \underset{\sim}{\mu}_j) = g(\mu_1, \cdots, \mu_k)$$

In general, computationally very difficult!

<u>Special case</u>: $d(\underset{\sim}{x}_i, \underset{\sim}{\mu}_j) = \|\underset{\sim}{x}_i - \underset{\sim}{\mu}_j\|^2$ = squared Euclidean norm

$\Rightarrow$ k-means

Re-express optimization problem:

$$\min_{\mu_1, \cdots, \mu_k} \min_{G_1, \cdots, G_k} \sum_{j=1}^{k} \sum_{i \in G_j} \|\underset{\sim}{x}_i - \underset{\sim}{\mu}_j\|^2$$

$$= \min_{G_1, \cdots, G_k} \min_{\mu_1, \cdots, \mu_k} \underbrace{\sum \sum \|\cdots\|^2}_{\text{same as above}}$$

minimized at mean of points in the cluster j

$$= \min_{G_1, \cdots, G_k} \underbrace{\sum_{j=1}^{k} \sum_{i \in G_j} \|\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}_{G_j}\|^2}_{\text{within group (cluster) sum of squares}}$$

where $\bar{x}_{G_j} = \dfrac{1}{\text{number of pts in } G_j} \sum_{i \in G_j} \underset{\sim}{x}_i$

<u>Comments</u>

① Still computationally very hard but good algorithms exist.

② In k-means clustering, shapes of clusters are all spheres (spheroids)

– can often transform variables (e.g. look at PC scores) s.t. shape assumption is not too severe.