Clustering: Votes Demo and Hierarchical Clustering

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

1 Votes demo

2 Hierarchical Clustering

Votes demo (clustering)

- Open R, read in data:

```
votes = dget(file = "C:/temp/votes.repub")
```

- Look at the votes data set by printing it to screen.
- We will group states according to their voting patterns; this dataset contains the % of republican votes in presidential elections.

Votes demo (clustering)

- remove Alaska and Hawaii; they become states in 1959 so there are not as many observations for them
- Let’s focus on 1916 & later elections, since the definitions of the two political parties has changed over time. Restrict the data to 1916 & later.

Votes demo (clustering)

- Call the kmeans function on the resulting data set. Start with "centers = 2" (two clusters). Instead of manually doing random restarts by repeatedly calling kmeans (like we did in lab), set “nstart = 200” so that the algorithm tries 200 different randomly generated sets of initial values.

- The resulting object (e.g. votesClust) has a vector votesClust$cluster giving the cluster number for each state.
- Print the names of all states in cluster 1. Same for cluster 2.

Votes demo (clustering)

- Look at the states in the 2 clusters. Do the clusters make sense?
- Increase the # of clusters one at a time until you get to a set of clusters that seems most "reasonable" to you

Votes demo (clustering)

We did not standardize the variables here; why is this reasonable?

Hierarchical Clustering

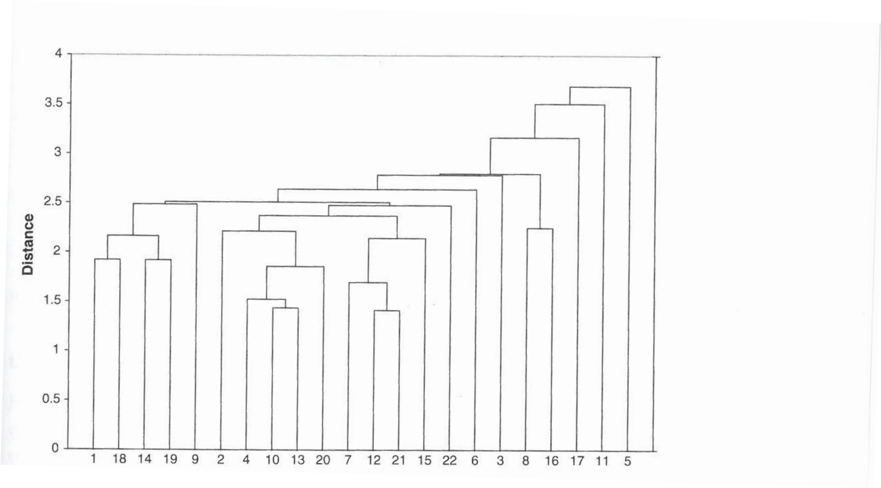
- **Hierarchical Clustering** is another distance-based clustering method.

- It avoids having to select K

- It provides an easily interpretable tree-based graphical representation of the data

Hierarchical Clustering

Example: “Dendrogram” (tree diagram) of the utilities (from SPB):



Hierarchical Clustering

Interpretation of a dendrogram:

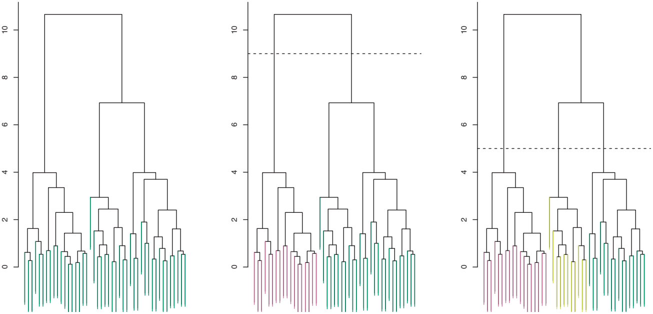
- Each leaf represents a single observation (here, one of the 22 utilities)
- As we move up the tree, some leaves begin to fuse into branches. These are observations that are similar to one another.
- As we move higher, branches fuse with each other or with leaves
- The lower in the tree a fusion occurs, the more similar the groups of observations are to one another
- Important: how close observations are on the x-axis does NOT indicate how similar they are
    - Ex:

Hierarchical Clustering

How to get clusters from a dendrogram?

Hierarchical Clustering

Example: Dendrogram of simulated data:



Hierarchical Clustering

- Any number of clusters between 1 and *n* can be obtained in this way

- So the height of the cut serves the same role as *K* in *K*-means: controls the # of clusters

- "Hierarchical" means that the clusters obtained by cutting at a particular height are **nested** within the clusters obtained by cutting at a greater height

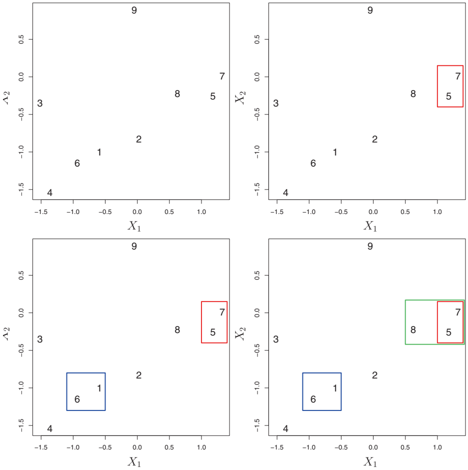
Hierarchical Clustering

Hierarchical Clustering Algorithm:

- Starting at the bottom of the dendrogram, each observation is in its own cluster
- The two closest observations (according to our selected distance measure) are **fused** to get $n - 1$ clusters
- Next the 2 clusters that are closest to each other are fused, giving $n - 2$ clusters
- This is repeated until there is only 1 cluster

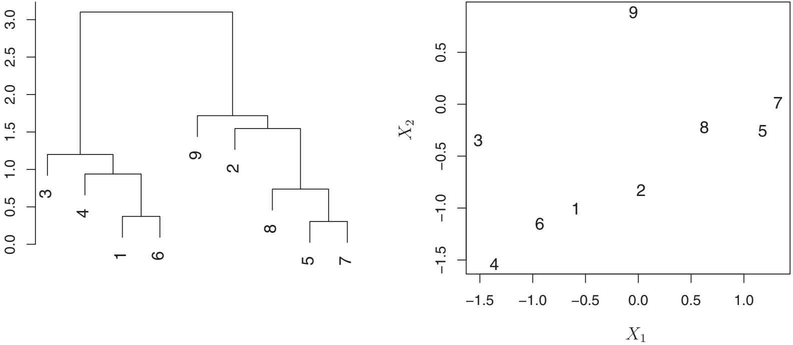
Hierarchical Clustering

A simple example with $p = 2$ continuous variables, $n = 9$, and using Euclidean distance:



Hierarchical Clustering

On left is the dendrogram for this example:



Hierarchical Clustering

Distance between clusters:

- But how did we decide to fuse the cluster $\{5, 7\}$ with the cluster $\{8\}$?