Model-Based Clustering

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Downsides of distance-based clustering

Downsides of distance-based clustering:

- Difficulty in capturing clusters of different sizes and shapes
  - k-means can't capture them. Only captures **round** clusters of equal sizes: <Draw picture of two side-by-side tall skinny clusters in $p = 2$ dimensions, that are somewhat close together on the x-axis. Then show that k-means will put the top half of both groups in one cluster, and the bottom half of both groups in the other cluster>
  - Hierarchical clustering: can only capture irregularly shaped clusters when "single linkage" is used to measure the distance between clusters, but single linkage is not preferred in practice because it tends to yield very imbalanced trees and long, trailing clusters

Downsides of distance-based clustering

Downsides of distance-based clustering:

- Sensitivity to the scaling of the variables <draw two scatterplots in $p = 2$ dimensions. The first has two round clusters next to each other horizontally. This is a dataset before standardization. The second plot shows what happens when standardization is applied to this dataset, which stretches out the clusters in the vertical direction. This is Figure 14.5 in Hastie, Tibshirani, and Friedman 2009. The two groups can be captured by k-means in the first plot, but in the second plot you have the problem described on the previous slide.>

- Hierarchical clustering tends to be unstable: slightly different datasets can yield very different trees

Model-Based Clustering

Model-based clustering:

- uses a probabilistic (statistical) model
- can capture long skinny clusters, and clusters of different sizes
- is not sensitive to the scaling of the variables

Model-Based Clustering

Model-based clustering:

- assumes that the observations x_i in a particular cluster k are drawn from some distribution
  - Example: assume $x_i \sim N(\mu_k, \sigma^2)$
- if we can learn the distributions associated with the different clusters then we can estimate to what cluster each observation belongs.
  - Same example: estimate the mean μ_k of each cluster, and σ^2. Then the observations x_i close to μ_k probably belong to cluster k

Model-Based Clustering

- Model-based methods (unlike heuristic clustering methods) yield the estimated **probability** that x_i belongs to each cluster k, instead of just an assignment to a cluster
- Captures our **uncertainty** in the cluster assignments!
- Although model-based methods do not have an explicit distance measure, they have an implicit one (discussed later)
- Sophisticated model-based methods are essentially able to **learn some aspects of the distance measure** from the data
- They can even allow the distance measure to be different for each cluster!

Model-Based Clustering

Although the method returns a probability, can always get an estimated encoder C by: Assigning each observation to the highest-probability cluster

I've developed model-based clustering methods for:

- **resolving performance problems in datacenters**
- **image segmentation** (distinguishing features of an image) for three-dimensional medical images

Model-Based Clustering Demo

- Download votes.repub, read into R, and take subset:

  votes = dget("C:/temp/votes.repub")
  votes = votes[-c(2,11), 16:31]

- Install the "mclust" package in R (Packages->Install Package->choose mirror->mclust). Read in:

  library(mclust)

- Fit the model to just 2 variables (to make it easy to visualize; in practice would use all variables):

  clust = Mclust(votes[,c(6,15)], G = 2)

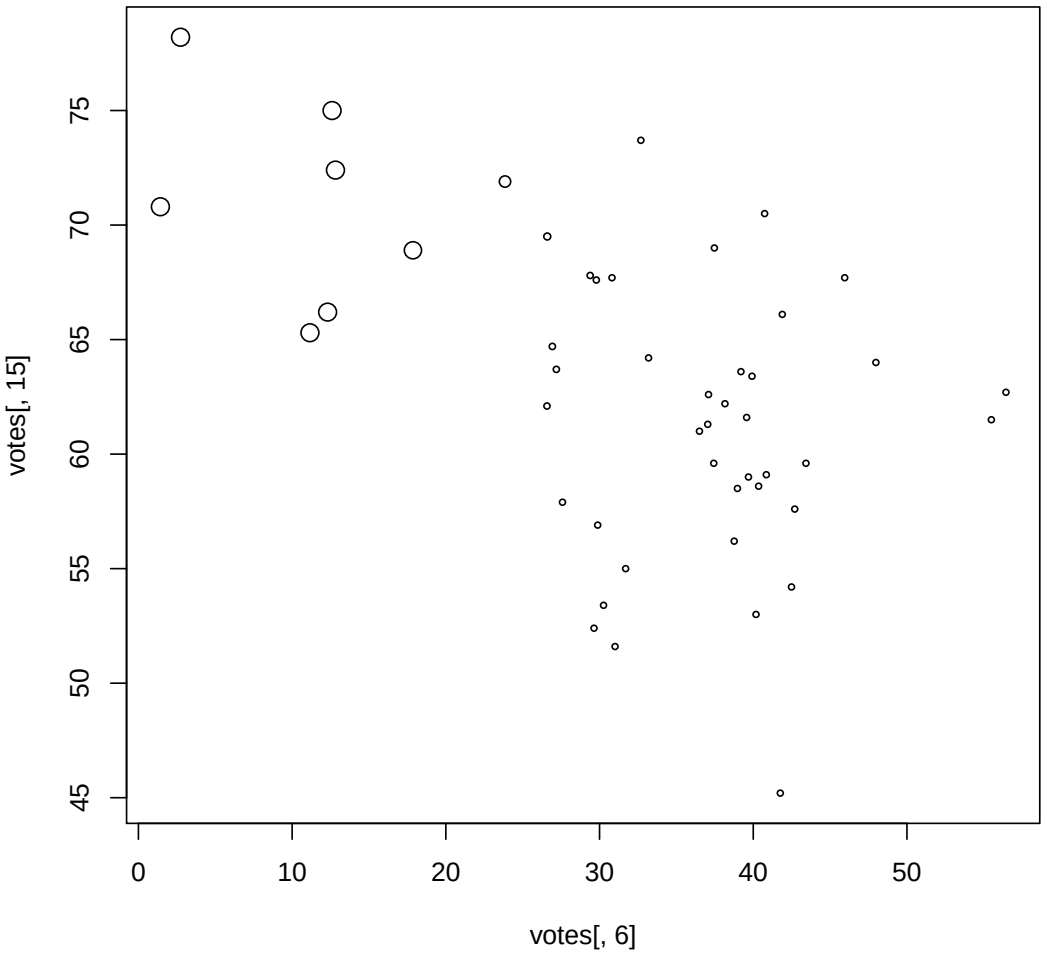
Model-Based Clustering Demo

- The matrix **clust$z** contains the estimated probability that each observation belongs to each cluster; what are its dimensions?
  48×2

- Plot the 2 variables in a scatterplot, where the size of the point corresponds to the probability that the observation belongs to cluster 1:

  plot(x = votes[,6], y = votes[,15], cex = (clust$z[,1]+.5))

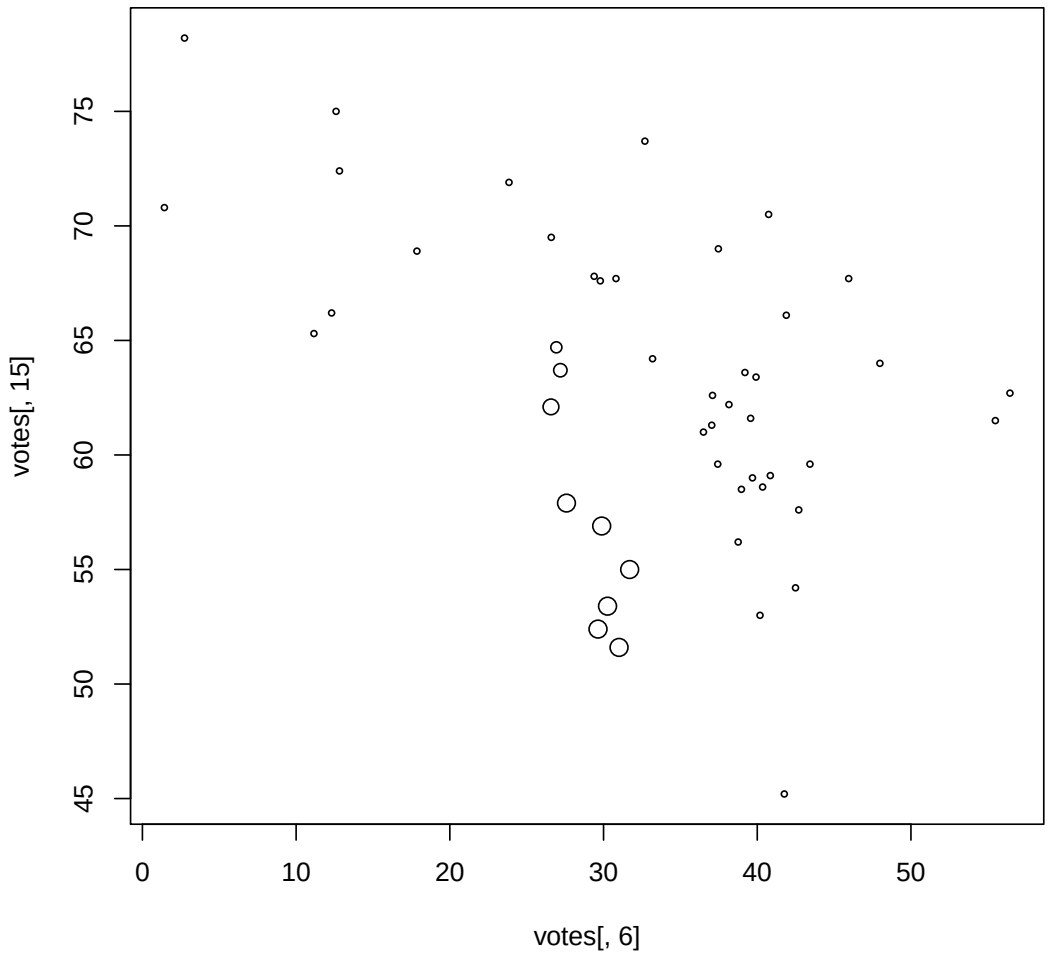
Model-Based Clustering Demo



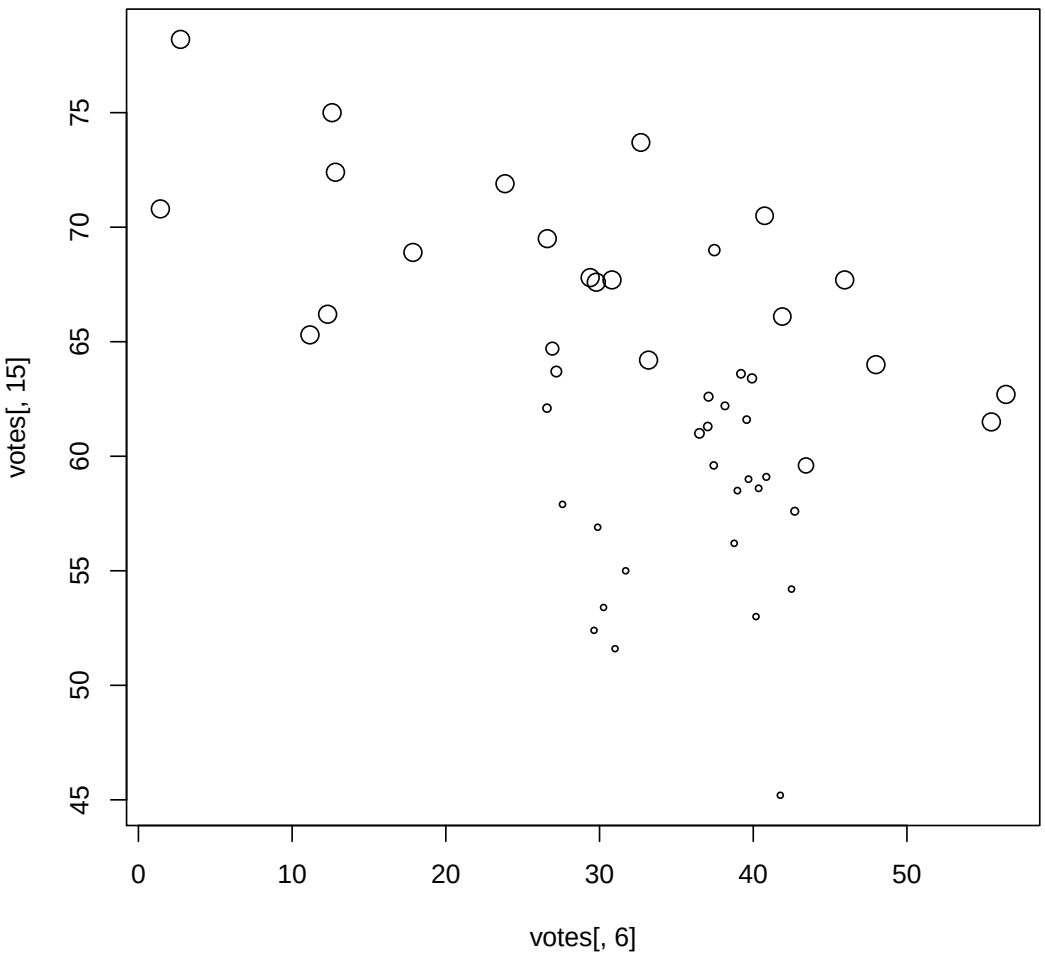
Interpret: One of the observations has probability that is close to 0.5. Others are sure what cluster they belong in.

Model-Based Clustering Demo

Increase the # clusters to 3, and create plots for the probability that each obs. belongs in each cluster k=1,2,3.



Model-Based Clustering Demo



Model-Based Clustering Demo

How else are the results different from what we would expect k-means to give?

The clusters are not round. K-means yields round clusters, because it is based on unweighted Euclidean distance.

Model-based clustering can capture non-round clusters, essentially because it learns the dissimilarity measure from the data, and separately for each cluster.

Model-Based Clustering

A simpler version of the model used in this demo assumes that, for unknown parameters $\{w_k, \mu_{kj}, \sigma^2_{kj} : \ k = 1, \ldots, K; \ j = 1, \ldots, p\}$:

1. Before seeing the data, each observation i has probability w_k of belonging to each cluster k
   - I.e., $Pr(C(i) = k) = w_k$

2. If observation i belongs to cluster k, then each variable j is sampled from a normal distribution with mean μ_{kj} and variance σ^2_{kj}
   - I.e., $X_{ij} \mid C(i) = k \ \sim \ \phi_{\mu_{kj}, \sigma^2_{kj}}(x_{ij})$
   - Or, $X_i \mid C(i) = k \ \sim \ \prod_{j=1}^{p} \phi_{\mu_{kj}, \sigma^2_{kj}}(x_{ij})$

Here $\sum_k w_k = 1$
remember that ϕ_{μ, σ^2} is the normal density with mean μ and variance σ^2

Model-Based Clustering

The parameters $\{w_k, \mu_{kj}, \sigma^2_{kj} : \ k = 1, \ldots, K; \ j = 1, \ldots, p\}$ are estimated by maximum likelihood (for details see Hastie, Tibshirani, and Friedman 2009, Sec. 6.8)

Then we can get the probability that observation i belongs to cluster k using Bayes' Theorem:

$$Pr(C(i) = k \mid X_i = x_i) = \frac{Pr(C(i) = k) f(x_i | C(i) = k)}{\sum_{k'=1}^{K} Pr(C(i) = k') f(x_i | C(i) = k')}$$

$$= \frac{w_k \prod_{j=1}^{p} \phi_{\mu_{kj}, \sigma^2_{kj}}(x_{ij})}{\sum_{k'=1}^{K} w_{k'} \prod_{j=1}^{p} \phi_{\mu_{k'j}, \sigma^2_{k'j}}(x_{ij})}$$

Model-Based Clustering

If we forced all the parameters σ_{kj}^2 to be the same (had a single variance parameter σ^2, we would get round clusters (like k-means). Why?

Why is it useful to allow them to be different?

The model-based clustering method implemented by “Mclust” actually allows nonzero covariances between the variables as well! This allows:

Model-Based Clustering