STAT 185 - Problem Set #7

NAME

December 3, 2019

Instructions

Hierarchical clustering: You may use the command *out <- hclust(DISS, method = "method")* where method may be "single", "average", or "complete" for the chosen linkage and *print(out)* displays the resulting dendrogram. The variable *DISS* is the dissimilarity or distance matrix for the data of interest. The command *cutree(out, k = K)* gives the associated K cluster solution.

k-medoids clustering: Use the command *out <- pam(data, K, metric = "euclidean")* to find a K cluster solution using k-medoids. The cluster labels are given in *out$clustering.*

k-means clustering: Use the command *out <- kmeans(data, centers = K)* to find a K cluster solution using k-means. The cluster labels are given in *out$cluster.*

Gaussian mixtures: Use the command *out <- Mclust(data, modelNames = name, G = K)* to find a K cluster solution using a Gaussian mixture model with K component and covariance structure given by name. To compare multiple models, use the command *out <- mclustBIC(data, G = , modelNames =).* Calling *plot(out)* shows the BIC curves for all methods and number of mixture components. Calling *summary(out)* gives the three best-performing methods and number of mixture component.

Spectral clustering: Use the command *out <- specc(data, centers = K)* to find a K cluster solution using spectral clustering. The algorithm defaults to a radial basis function with adaptively determined width parameter given in *out@kernelf@kpar.* The cluster labels are given in *out@.Data.*

Problems

1. (Blum 7.1) Construct examples where using distances instead of distance squared gives bad results for Gaussian densities. For example, pick samples from two 1-dimensional unit variance Gaussians, with their centers 10 units apart. Cluster these samples by trial and error into two clusters, first according to k-means and then according to the k-median criteria. The k-means clustering should essentially yield the centers of the Gaussians as cluster centers. What cluster centers do you get when you use the k-median criterion?

2. (Hastie 14.5) Generate data with three features, with 30 data points in each of three classes as follows:

$$\begin{aligned}
\theta_1 &= U(-\pi/8, \pi/8)\\
\phi_1 &= U(0, 2\pi)\\
x_1 &= \sin(\theta_1)\cos(\phi_1) + W_{11}\\
y_1 &= \sin(\theta_1)\sin(\phi_1) + W_{12}\\
z_1 &= \cos(\theta_1) + W_{13}
\end{aligned}$$

$$\begin{aligned}
\theta_2 &= U(\pi/4, 3\pi/4)\\
\phi_2 &= U(-\pi/4, \pi/4)\\
x_2 &= \sin(\theta_2)\cos(\phi_2) + W_{21}\\
y_2 &= \sin(\theta_2)\sin(\phi_2) + W_{22}\\
z_2 &= \cos(\theta_2) + W_{23}
\end{aligned}$$

$$\begin{aligned}
\theta_3 &= U(\pi/4, 3\pi/4)\\
\phi_3 &= U(\pi/4, 3\pi/4)\\
x_3 &= \sin(\theta_3)\cos(\phi_1) + W_{31}\\
y_3 &= \sin(\theta_3)\sin(\phi_1) + W_{32}\\
z_3 &= \cos(\theta_3) + W_{33}
\end{aligned}$$

Here $U(a,b)$ indicates a uniform variate on the range $[a,b]$ and W_{jk} are independent normal variates with standard deviation 0.6. The data lie near the surface of a sphere in three clusters centered at $(1,0,0)^T$, $(0,1,0)^T$ and $(0,0,1)^T$. Carry out a K-means, K-medoid, and Model-based clustering of the same data and compare the results. In all three cases, assume 3 components. For the model based clustering use *modelNames = "EII"* fir a Gaussian mixture models with covariance $\sigma^2 I$ in all three clusters.

3. (Hastie 14.2) Consider a mixture model density in p-dimensional space,

$$g(x) = \sum_{k=1}^{K} \pi_k g_k(x),$$

where $g_k(x) = \mathcal{N}(\mu_k, \sigma^2\mathbf{I})$ and $\pi_k \geq 0$ for all k with $\sum_k \pi_k = 1$. Here $\{\mu_k, \pi_k\}_{k=1}^K$ and σ^2 are the unknown parameters. Suppose we have data $x_1, \ldots, x_N \sim g(x)$ and we wish to fit the mixture model.

   a. Write down the log-likelihood of the data.

   b. Derive an EM algorithm for computing the maximum likelihood estimates of the unknown parameters.

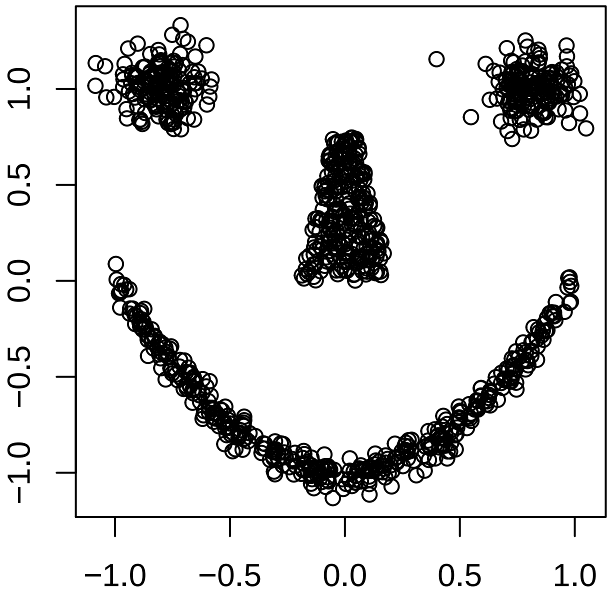
   c. Show that if σ has a known value in the mixture model and we take $\sigma \to 0$, then in a sense this EM algorithm conincides with K-means clustering.

4. (Izenman 12.3) The variable *primate.scapulae* consists of measurements of indices and angles on the scapulae (shoulder bones) of five genera of adult primates representing Hominoidae: gibbons (Hylobates), orangutangs (Pongo), chimpanzees (Pan), gorillas (Gorilla), and man (Homo). The variables of interest are 5 indices (AD.BD, AD.CD, EA.CD, Dx.CD, and SH.ACR) and 2 angles (EAD, β). The *gamma* variable is missing for some data so we will ignore it. Of the 105 measurements on each variable, 16 were from Hylobates, 15 from Pongo, 20 from Pan, 14 from Gorilla, and 40 from Homo.

   a. Cluster the *primate.scapulae* data using single-linkage, average-linkage, and complete-linkage agglomerative hierarchical clustering using Euclidean distance. Print the dendrogram for each method.

   b. Find the five-cluster solutions for all three methods which allows comparision with the true primate classifications which are stored in the variable *primate.scapulae$classdigit*. Using the 5-cluster solution, find the misclassification rate for all three methods. Show that the lowest rate occurs for the complete-linkage method and the hightest for the single-linkage method.

5. A plot of the data contained in the matrix *smile* is comprised of four pieces (two eyes, a nose, and a mouth) shown below.



a. Plot the BIC curves vs. number of clusters for each of the default Gaussian mixture models in the MClust package. Which method gives the best BIC? Plot the data with points colored by cluster in the optimal solution. Do the BIC estimates effectively estimate the correct number of components? Explain.

b. Plot the *smile* data coloring points by the cluster ID, assuming 4 clusters, given by the best model-based clustering by BIC.

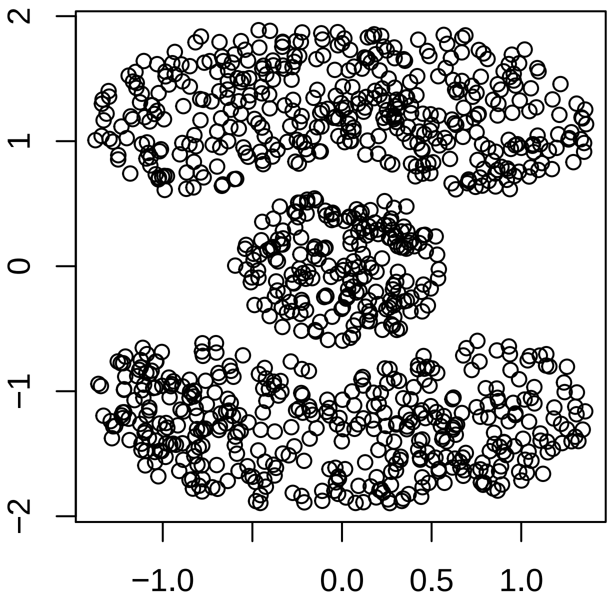
c. Plot the eigenvalues of the graph Laplacian matrix for spectral clustering using a radial basis function with bandwidth $\sigma = 1/150$. Do the eigenvalues effectively estimate the number of components? Explain.

d. Plot the *smile* data coloring points by cluster ID, assuming 4 clusters, given by spectral clustering.

e. Show the clustering results, assuming 4 clusters, using single-linkage, average-linkage, and complete-linkage.

f. Compare the results of all of the clustering methods used in this problem.

6. Repeat problem 6 using the data contained in the variable *cassini* comprised of three components shown below.



a. Plot the BIC curves vs. number of clusters for each of the default Gaussian mixture models in the MClust package. Which method gives the best BIC? Plot the data with points colored by cluster in the optimal solution. Do the BIC estimates effectively estimate the correct number of components? Explain.

b. Plot the *smile* data coloring points by the cluster ID, assuming 3 clusters, given by the best model-based clustering by BIC.

c. Plot the eigenvalues of the graph Laplacian matrix for spectral clustering using a radial basis function with bandwidth $\sigma = 1/800$. Do the eigenvalues effectively estimate the number of components? Explain.

d. Plot the *smile* data coloring points by cluster ID, assuming 3 clusters, given by spectral clustering.

e. Show the clustering results, assuming 3 clusters, using single-linkage, average-linkage, and complete-linkage.

f. Compare the results of all of the clustering methods used in this problem.