CPH 636 — Spring 2009 — Dr. Charnigo

Lecture 12

Clustering

Introduction. Clustering is an unsupervised learning operation in which we separate the data into groups or "clusters" of similar observations according to their values on continuous variables $X_1, \ldots, X_j$. Two common approaches to clustering are "K-means clustering" and "hierarchical clustering".

In "K-means clustering" we specify that there will be K clusters. (We can choose K on our own, or we can use diagnostic plots from Fernandez's DISJCLUS macro.) We place each observation in the cluster to which it is closest in j-dimensional space. Details will be provided in this lecture.

In "hierarchical clustering" we do not fix the number of clusters. We begin with a single cluster containing all observations, we break the single cluster into two clusters, we break one of the two clusters so that we have three, we break one of the three clusters so that we have four, and so forth, until each observation is its own cluster. A visual representation of this hierarchy is called a dendogram. Figure 2B of

`http://clincancerres.aacrjournals.org/cgi/content/full/10/18/6143`

provides an illustration. The left dendogram separates human beings into clusters, while the upper dendogram separates genetic markers into clusters. While one can get a specific number of clusters of human beings by drawing a vertical line through the left dendogram, the main value of the left dendogram is that it shows how similar any two human beings are: simply start at the corresponding rows of the array and see how far left you must go for the dendogram branches to merge. We will not pursue hierarchical clustering further since Fernandez has not created a macro for it.

Notation for K-means clustering. We will assume that the continuous variables $X_1, \ldots, X_j$ have been standardized. (Fernandez's DISJCLUS macro takes care of the standardization for us automatically.)

Let

$$\mathbf{x}_i := (x_{1,i}, \ldots, x_{j,i})^T$$

denote the vector of values on $X_1, \ldots, X_j$ for subject i in the data set. Let

$$\mathbf{c}_1 := (c_{1,1}, \ldots, c_{j,1})^T$$

denote a vector of numbers representing the center of the first cluster in j-dimensional space. Let

$$\mathbf{c}_2 := (c_{1,2}, \ldots, c_{j,2})^T$$

through

$$\mathbf{c}_K := (c_{1,K}, \ldots, c_{j,K})^T$$

denote vectors of numbers representing the centers of the remaining clusters.

The distance between the observation for subject i and the center of cluster 1 is

$$d(\mathbf{x}_i, \mathbf{c}_1) := \sqrt{(x_{1,i} - c_{1,1})^2 + (x_{2,i} - c_{2,1})^2 + \cdots + (x_{j,i} - c_{j,1})^2},$$

the distance between the observation for subject i and the center of cluster 2 is

$$d(\mathbf{x}_i, \mathbf{c}_2) := \sqrt{(x_{1,i} - c_{1,2})^2 + (x_{2,i} - c_{2,2})^2 + \cdots + (x_{j,i} - c_{j,2})^2},$$

and so forth.

Procedure for K-means clustering. We must specify K when we run Fernandez's DISJCLUS macro. (Thus, if we are going to use the macro's diagnostic output to choose K, we will have to run the macro twice.) Given K, the macro automatically determines "initial values" for the elements of $\mathbf{c}_1, \ldots, \mathbf{c}_K$. Subject i is provisionally assigned to cluster 1 if

$$d(\mathbf{x}_i, \mathbf{c}_1) = \min\{d(\mathbf{x}_i, \mathbf{c}_1), d(\mathbf{x}_i, \mathbf{c}_2), \ldots, d(\mathbf{x}_i, \mathbf{c}_K)\},$$

to cluster 2 if

$$d(\mathbf{x}_i, \mathbf{c}_2) = \min\{d(\mathbf{x}_i, \mathbf{c}_1), d(\mathbf{x}_i, \mathbf{c}_2), \ldots, d(\mathbf{x}_i, \mathbf{c}_K)\},$$

and so forth. That is, each observation is provisionally assigned to the cluster to which it is closest.

After these provisional assignments are made, the macro determines new values for the elements of $\mathbf{c}_1, \ldots, \mathbf{c}_K$. The new values for the elements of $\mathbf{c}_1$ are the means of X_1 through X_j among all subjects provisionally assigned to cluster 1, the new values for the elements of $\mathbf{c}_2$ are the means of X_1 through X_j among all subjects provisionally assigned to cluster 2, and so forth.

After $\mathbf{c}_1$ through $\mathbf{c}_k$ have been updated, each observation is (re)assigned to the cluster to which it is closest. Then $\mathbf{c}_1$ through $\mathbf{c}_k$ are updated a second time, following which each observation is (re)assigned to the cluster to which it is closest. This iterative process continues until the cluster assignments stabilize.

Illustrative example. I will discuss some practical aspects of K-means clustering in the course of presenting an illustrative example. I will be referring to the output files {GFLIB.DIABET140.doc} and {GFLIB.DIABET141.doc}, which are available from my web page. The former output file was obtained by running the DISJCLUS macro with the exploratory analysis option on, while the latter output file was obtained by running the DISJCLUS macro with the exploratory analysis option off.

I applied the DISJCLUS macro to the first diabetes data set supplied by Fernandez, which is available in {diabet1.sas7bdat}. We previously considered this data set when discussing discriminant analysis in Lecture 7. However, I have now discarded diabetic status (normal, overt diabetic, chemical diabetic), which we had regarded as the response variable in Lecture 7. Recall that the other variables, X_1 through X_5, represented relative weight, fasting plasma glucose level, test plasma glucose, plasma insulin during test, and steady-state plasma glucose level.

Although you do not have a written assignment on K-means clustering, instructions for using Fernandez's DISJCLUS macro are available in {SASMacroInstr8.txt}.

Discussion question. What reason would there be to apply clustering, an unsupervised learning operation, to the explanatory variables in a data set that had a response variable?

Cluster means and standard deviations. At the bottom of page 1 of {GFLIB.DIABET140.doc} are cluster means and standard deviations for the standardized versions of $X_1, \ldots, X_5$. (Note that I decided to form three clusters.) For example, the mean and standard deviation of (standardized) X_1 within cluster 1 are -0.536 and 0.723. The within standard deviation for X_1, reported at the top of page 1, is the square root of the pooled variance,

$$\sqrt{\frac{(69-1)\times(0.723)^2+(45-1)\times(0.631)^2+(27-1)\times(0.982)^2}{69+45+27-3}} = 0.752.$$

Above, the $69, 45, 27$ represent the numbers of subjects assigned to the three clusters (see page 2). The R^2 value of 0.442 represents the fraction of variability in X_1 accounted for by cluster membership. The $R^2/(1-R^2)$ value of 0.791 reflects the ratio of between-cluster variability in X_1 to within-cluster variability in X_1.

Variable selection. Page 6 presents the results of applying a backward elimination procedure to investigate whether clustering may be based on a reduced variable set. Although X_5 is recommended for elimination, the rest of the results in {GFLIB.DIABET140.doc} assume that $X_1, \ldots, X_5$ have been used for clustering rather than only $X_1, \ldots, X_4$. (Thus, if we really wanted to use only $X_1, \ldots, X_4$, we would need to run the macro again.)

Scatter plots. Page 7 contains scatter plots of X_4 against X_5, X_2 against X_5, and so forth. Plotting symbols of "1", "2", and "3" identify the clusters to which the subjects have been assigned. Interestingly, these plots depict the unstandardized versions of $X_1, \ldots, X_5$ rather than the standardized versions. We see that X_1 and X_2 nicely separate the clusters: subjects with large X_2 values are in cluster 3, subjects with small X_2 values and large X_1 values are in cluster 2, subjects with small X_2 values and small X_1 values are in cluster 1.

CCC plot. Page 8 is a diagnostic plot that can aid in the choice of K if we are not satisfied with the choice we made when we ran the macro. The "cubic clustering criterion" (CCC) is plotted against the number of clusters. Fernandez says that CCC values greater than about 2 suggest effective clustering. In this example, the CCC is almost 2 when there are three clusters and is not that large again until there are twelve clusters, which may be many more than we would want to have.

PSF and PST plot. Page 9 is another diagnostic plot that can aid in the choice of K. The "pseudo F statistic" (PSF) and "pseudo T^2 statistic" (PST) are plotted against the number of clusters. Fernandez says that the "elbow points" in this plot can suggest an appropriate number of clusters. The PSF suggests taking $K = 8$, while the PST suggests taking $K = 5$.

Cluster means and standard deviations. Now let us examine {GFLIB.DIABET141.doc}. Page 1 is identical to page 1 of the first output file since I did not change K or the variables used when I ran the macro with the exploratory analysis option off.

Assessing multivariate normality. There are several pages of output in {GFLIB.DIABET141.doc} dedicated to assessing whether $X_1, \ldots, X_5$ have a multivariate normal distribution within each cluster. Pages 5, 10, 15 contain quantile-quantile plots as well as p-values for multivariate skewness and kurtosis within the three clusters. (If the p-values are illegible on these pages, they can also be found on pages 3, 8, 13.)

Cluster assignments. Page 17 reports the numbers of subjects assigned to each cluster. In this example, 69, 45, and 27 subjects were assigned to clusters 1, 2, and 3.

Page 21 provides partial results for a canonical discriminant analysis with cluster assignments treated as if they were values of a response variable. The canonical variables C_1 and C_2 are not explicitly defined, although we are provided with pairwise correlations between $X_1, \ldots, X_5$ and C_1, C_2. For instance, the correlation between X_1 and C_2 is 0.867.

Page 22 reports the means of C_1 and C_2 within the three clusters. Pages 23 through 26 explicitly identify which subjects are assigned to each cluster and provide each subject's values of C_1, C_2.

Graphical displays. Pages 27 and 28 provide box plots of C_1 and C_2 within each cluster. We see that C_1 completely separates cluster 3 from the other two, while C_2 nearly separates cluster 1 from cluster 2.

Page 29 is a scatter plot of C_2 against C_1 with plotting symbols of "1","2","3" used to identify the clusters to which subjects have been assigned. The arrows labeled $X_1, \ldots, X_5$ reflect the correlations between $X_1, \ldots, X_5$ and C_1, C_2. For instance, X_1 has a slight negative correlation with C_1 (a small leftward component for the arrow) but a strong positive correlation with C_2 (a large upward component for the arrow). The "+" symbols represent the locations of the cluster means for C_1 and C_2.