Statistical Clustering

It clearly would not be proper to combine the heights of men belonging to two dissimilar races, in the expectation that the compound results would be governed by the same set of constants….-Francis Galton

Introduction 3.0

The chapter on classification deals with methods to assign an observation to one of a finite number of classes. The classes are defined *a priori* and serve as a training set in the learning effort. Classification is valuable in being able to identify the class membership of a new observation based on a distance measure or some other criterion based on probabilistic analysis of the training data. The methodology is appropriate to assign new observations to previously established classes. It requires significant manual effort to create a comprehensive list of classes and populate the classes with examples. On the other hand, statistical clustering is a method that attempts to dissemble a heterogeneous set of observations into natural homogeneous groups such that within group homogeneity is small. Clustering can be viewed as an exploratory data analytic technique. Large e-commerce sites may

Notes:

want to discern differences in a large set of on-line visitors to understand their propensity to buy is one example of clustering.

The goal of statistical clustering is data dissection. Statistical clustering does not rely on a set of pre defined classes and in a sense can be viewed as an *ad hoc* procedure. Very little if any of effort is required to initiate clustering. Cluster analysis of data tends to unravel groups that are hitherto unknown. If clustering is performed over sets of data assembled periodically, the data sets break away into natural groupings and certain groups may highlight newly emerging data segments. For example, if one is studying the voting trends in a district, clustering may reveal a groupings of voters concerned about specific issues due to changes in social, economic, or security environments.

A distinguishing feature of clustering is that it is not for the large part based on probabilistic models. More formally, the process of grouping a set of physical or abstract entities into *homogenous* or *similar* classes is called clustering. A cluster is a collection of data entities that are *similar* to one another within the same cluster

Notes:

and quite dissimilar to objects or entities in other clusters. There are mainly three approaches to statistical clustering known as *hierarchical clustering*, clustering based on *partition methods* and finally model based clustering. Our focus is on these methods in this chapter.

Section 3.1 Hierarchical clustering

Hierarchical clustering methods involve a series of iterations combining or dividing a group of observations. Methods involving successive mergers of observations are known as agglomerative procedures, and those that involve a series of divisions are known as divisive. Agglomerative clustering start by treating each individual observation in the data set as a cluster. Then the most similar observations are combined into groups and in the subsequent steps, the groups are further fused based on their similarity. Divisive clustering takes the opposite approach. At the outset it treats the entire data set as one giant cluster. Subsequently the data set is broken into two subsets such that observations in one subset are far from the observations in the other. The subsets are further divided into component dissimilar

Notes:

clusters and the process continues until there are as many sub-groups as the number of observations. The results of agglomerative and divisive clustering are displayed visually by a tool known as a *dendrogram*. A dendrogram is a 'tree-like' diagram that summaries the process of clustering. Similar cases are joined by links whose position in the diagram is determined by the degree of similarity between the observations.

Section 3.2 Agglomerative clustering

In agglomerative clustering, each record is treated as a cluster. If there are *n* records, there are *n* clusters to begin with, each containing one record. The next is a partition into $n-1$clusters, and the next into $n-2$, and so on…..until all samples form one cluster. Clearly at the k^{th} level, the number of clusters is equal to $n-k+1$. Algorithmically, agglomerative clustering may be laid out in the following steps.

- Start with an initial set of *N* clusters, each cluster consisting of a single object. The pair-wise distances between objects in the singleton clusters are arranged in a

Notes:

matrix form. The matrix denoted by *C* is a symmetric matrix of order *N* x *N*.

$$C = \begin{bmatrix} d_{11} & d_{12} & \cdots\cdot & d_{1N} \\ d_{21} & d_{22} & \cdots\cdot & d_{2N} \\ . & . & \cdots\cdot & . \\ d_{N1} & d_{N2} & \cdots\cdot & d_{NN} \end{bmatrix} \qquad (3.1)$$

Each $d_{ij}, i = 1,2,....N, j = 1,2,.....,N$ is the Euclidean distance between objects x_i and x_j.

- Scan the matrix C for the nearest pair of objects. Find the closest pairs of objects and merge them into one cluster.
- Then delete the rows and columns corresponding clusters that were merged and adding a row and column giving the pair-wise distances between the newly formed cluster and the remaining clusters.
- Steps two and three are repeated $N-1$ times till objects combine to form one cluster.

As a part of book keeping a log of merged clusters and their corresponding levels is maintained.

Notes:

Procedurally, Agglomerative clustering procedure is described in the following steps:

Procedure: Basic agglomerative Clustering

Loop:

1. Let $\hat{C} = n$ and $\{ X_i \} = \{ x_i \}, i = 1,2,3,....n.$
2. If $\hat{C} \le c$, stop
3. Find the nearest pair of distinct clusters, say X_i and X_j.
4. Merge X_i and X_j, and delete X_j, and decrement $\hat{C}$ by one.
5. Go to Loop

Note that in step 3. We need to find the two nearest X_i and X_j.

Consider the matrix of distances

$$C = C_{ij} = \begin{bmatrix} 0 & 6 & 2 & 5 & 8 & 10 \\ 6 & 0 & 6 & 4 & 9 & 9 \\ 2 & 6 & 0 & 12 & 4 & 6 \\ 5 & 4 & 12 & 0 & 5 & 1 \\ 8 & 9 & 4 & 5 & 0 & 8 \\ 10 & 9 & 6 & 1 & 8 & 0 \end{bmatrix} \qquad (3.2)$$

Notes:

The matrix ***C*** is capturing the pair wise distances between objects. The objects might be a litter of mice and the entries (observations) the difference in weight in grams between any pair of mice in the litter. Scanning the matrix of distances note that objects 6 and 4 are the closest. Combining objects 6 and 4, the distances between (64) and 1,2,3,5 are:

$$\begin{aligned} d_{(64)1} &= \min(d_{61}, d_{41}) = \min(10,5) = 5 \\ d_{(64)2} &= \min(d_{62}, d_{42}) = \min(9,4) = 4 \\ d_{(64)3} &= \min(d_{63}, d_{43}) = \min(9,12) = 9 \\ d_{(64)5} &= \min(d_{65}, d_{45}) = \min(8,5) = 5 \end{aligned} \qquad (3.3)$$

Using the distances, the new matrix of distances is

$$\begin{array}{c|ccccc} & (64) & 1 & 2 & 3 & 5 \\ \hline (64) & 0 & & & & \\ 1 & 5 & 0 & & & \\ 2 & 4 & 6 & 0 & & \\ 3 & 6 & \boxed{2} & 6 & 0 & \\ 5 & 5 & 8 & 9 & 4 & 0 \end{array}$$

Figure 3.1. The new configuration of distances after the first step in the clustering process

From the matrix it is clear that objects 3 and 1 are the next set of objects to be combined. At this stage in the agglomerative process, the sub groups are, (64), (31),

Notes:

2, and 5. The distances between (64) and (31) is

$$\begin{aligned}
d_{(64),(31)} &= \min\left(d_{(64)3}, d_{(64)1}\right) = (6,5) = 5 \\
d_{(64),2} &= \min\left(d_{62}, d_{42}\right) = (9,4) = 5 \\
d_{(64),5} &= \min\left(d_{65}, d_{45}\right) = (8,5) = 5 \\
d_{(31),2} &= \min\left(d_{32}, d_{12}\right) = (6,6) = 6 \\
d_{(31),5} &= \min\left(d_{35}, d_{15}\right) = (4,8) = 4
\end{aligned} \tag{3.4}$$

The new matrix of distances is

$$\begin{array}{c|cccc}
 & (64) & (31) & 2 & 5 \\
\hline
(64) & 0 & & & \\
(31) & 5 & 0 & & \\
2 & \boxed{4} & 6 & 0 & \\
5 & 5 & 4 & 9 & 0
\end{array}$$

Figure 3.2. Further clustering of objects in the sub-sequent step of clustering

From the matrix it is quite clear that (64) combines with 2 and 5 combines with (31). Finally, (642) and (315) combine in the last stage. The dendrogram upon clustering is given in Figure 3.3.

Notes:

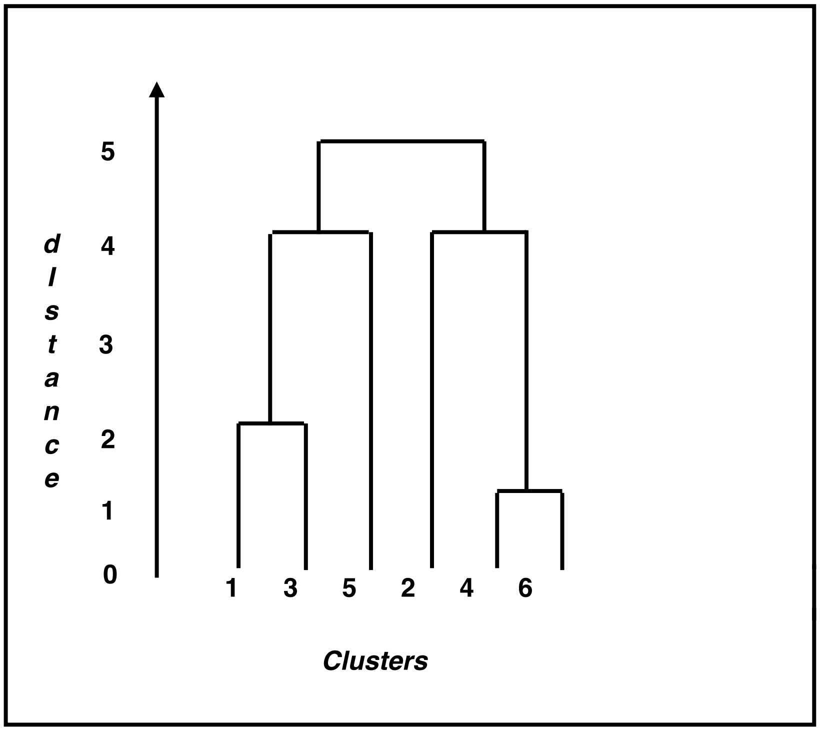

Figure *3.3. Dendrogram of the matrix C of distances given above.*

Section 3.3 Deficiencies of hierarchical methods

As with many other clustering schemes, clusters formed by agglomerative methods may be erroneous as linkages built early in the clustering process may be incorrect. The procedures tend to be sensitive to outliers. The consistency of statistical clustering may be established by adding some small errors to the observations and

Notes:

comparing the result to the outcome from the original set of unperturbed observations.

Agglomerative hierarchical clustering is discussed in Anderberg (1973), Sneath and Sokal (1973), Hartigan (1975), Everitt (1980), and Spath (1980). A very good introductory treatment is in Massart and Kaufman (1983). A serious practitioner of hierarchical cluster analysis should study the Monte Carlo results of Milligan (1980), Milligan and Cooper (1985), and Cooper and Milligan (1988). Other notable references on hierarchical clustering include Hartigan (1977, pp. 60 - 68; 1981), Wong (1982), Wong and Schaack (1982), and Wong and Lane (1983).

Section 3.4 Clustering by Partition Methods

The k -means algorithm is a very popular clustering method. It requires partitioning the data set into k preliminary partitions. It iteratively assigns observations into one of k clusters using a distance metric under a minimization criterion. It combines an effective method for finding initial clusters with a standard iterative algorithm for

Notes:

minimizing the sum of squared distances from the cluster means. The result is an efficient procedure for disjoint clustering of large data sets. A set of observations called *cluster centroids* is selected by computing the arithmetic mean of each partition. Each observation is assigned to the nearest centroid to form temporary clusters. The centroids are then updated by re-computing the centroids of the temporary clusters, and the process is repeated until no further changes occur in the clusters.

Similar techniques are described in most references on clustering (Anderberg 1973; Hartigan 1975; Everitt 1980; Spath 1980). The clustering is done on the basis of Euclidean distances computed from one or more numeric variables. Observations that are very close to each other are usually assigned to the same cluster, while observations that are far apart are in different clusters.

k-means clustering procedure is described in the following steps:

Procedure: Basic k-means clustering

Loop:

1. Proceed through the list of items, assigning an item to the cluster whose cluster center (centroid) is nearest. The measure of *"nearness"* is the Euclidian distance, computed with either standardized[†] or un-standardized

[†] standardized data is also referred to by some authors as normalized data. One method of standardization involves squeezing the range of a variable between 0 and 1. This is usually achieved by dividing each observation by the largest value in the data set.

observations. Recalculate the cluster center for the cluster receiving the new item and for the cluster losing the item.

2. Go to loop, until no more assignments take place.

Alternately, rather than starting with a partition of all items into k preliminary groups in step 1, we could specify k initial cluster centers and then proceed to step 2.

Iterative assignment of items to clusters is processed such that the within cluster variance is minimized and between cluster variance is maximized. In other words we are minimizing the squares of differences between each observation and a cluster center across clusters.

Example 3.1. Applying the K-means algorithm to fisher iris data.

```
data iris;
      title 'Fisher (1936) Iris Data';
      input SepalLength SepalWidth PetalLength PetalWidth Species @@;
      format Species specname.;
      label SepalLength='Sepal Length in mm.'
            SepalWidth ='Sepal Width in mm.'
            PetalLength='Petal Length in mm.'
            PetalWidth ='Petal Width in mm.';
      symbol = put(species, specname10.);
      datalines;
   50 33 14 02 1
   64 28 56 22 3
   65 28 46 15 2
   67 31 56 24 3
   .  .  .  .  .
   .  .  .  .  .
   52 35 15 02 1
   53 37 15 02 1
   ;

   proc fastclus data=iris maxc=2 maxiter=10 out=clus;
      var SepalLength SepalWidth PetalLength PetalWidth;
   run;
```

Section 3.5 Some practical issues, deficiencies and helpful hints associated with partition methods

Before performing cluster analysis on numerical data, it is necessary to consider scaling or transforming the variables since variables with large variances tend to have more effect on the resulting clusters than those with small variances. The data may be standardized such that variables have mean zero and variance equal to equal to 1. However standardization is not always desirable. Milligan and Cooper (1987) report a Monte Carlo study on various methods of variable standardization. Outliers or any suspicious observations should be removed before clustering.

Nonlinear transformations of the variables may change the number of population clusters and therefore should be applied judiciously. For most applications, the variables should be transformed so that equal differences are of equal practical importance. Non linear transformations are suitable for numeric data only.

If two or more initial points assigned as initial centroids lie in the same cluster, their resulting clusters may be poorly differentiated.

Notes:

The selection of k clusters at the outset forces the observations to be accommodated into the k clusters possibly leading to spurious results.

The presence of outliers in the data may produce clusters with highly variable data points.

Section 3.6 Gaussian Mixture Models

Heretofore, the clustering problem consisted of splitting a given data set into component clusters based on a distance measure subject to some optimization criterion such as minimizing sum of squares. Absent in this framework is a comprehension of an underlying stochastic process governing the observed data. To bring to bear the influence of the stochastic process, we model the clustering problem as a mixture of Gaussian densities. This approach to clustering is often referred to as *model based clustering*. It is model based in that there is an underlying component probability density function associated with each cluster. The idea of density based clustering is sound in that it recognizes and quantifies the random process producing the data. The probability distribution of the data is modeled as a

Notes:

linear combination of k densities and the method of maximum likelihood is used to estimate the component density parameters. The simplicity of making final assignment of cases to classes that it is most likely to have generated the observation makes the approach intuitive and appealing. Mathematically the probability density of the observed datum is expressed as a linear combination given by:

$$f\left(X_i \mid \underset{\sim}{\theta}\right) = \sum_{j=1}^{c} f\left(X_i \mid c_j, \theta_j\right)\pi_j \tag{3.5}$$

Where $\underset{\sim}{\theta}$ is a vector of unknown parameters of the mixture density, π_j, $j = 1,2,,....,c$ are known as mixing proportions or prior probabilities of the *j* clusters and c_j denotes the j^{th} class. Under this formulation, each $f\left(x_i \mid c_j, \theta_j\right)$ is modeled by a specific probability density function. The parameter estimates are obtained by maximum likelihood and are given as:

$$\pi_j = \frac{\sum_{j=1}^{c} f\left(c_j \mid x_i, \theta_j\right)}{n} \tag{3.6}$$

$$\mu_j = \frac{\sum_{i=1}^{n} f\left(c_j \mid x_i, \mu_j\right)x_i}{\sum_{i=1}^{n} f\left(c_j \mid x_i, \mu_j\right)} \tag{3.7}$$

Similarly a maximum likelihood estimate of the variance of a component density is given by;

$$\sigma_j^2 = \frac{\sum_{i=1}^{n} f\left(c_j \mid x_i\right)\left\|x_i - \mu_j\right\|^2}{\sum_{i=1}^{n} f\left(c_j \mid x_i\right)} \tag{3.8}$$

Notes:

The solution appears to be appealing as an estimate of the mean μ_j is a weighted average of the observations. The weight of sample x_i is simply an estimate of the likelihood of its membership in class c_j. We notice however that the posterior densities themselves are functions of μ_j. Thus the posterior function itself may be estimated by

$$\hat{f}\left(c_j \mid x_i, \hat{\mu}_j\right) = \frac{\hat{f}\left(x_i | c_j, \hat{\mu}_j\right)\hat{\pi}_j}{\sum_{j=1}^{c} \hat{f}\left(x_i \mid c_j, \hat{\mu}_j\right)\hat{\pi}_j} \tag{3.9}$$

Section 3.7 Final remarks

Clustering is a challenging field of research in data mining. As we saw there is a vast amount of literature on clustering dating back to several decades. The field is rife with new techniques. The success or failure of these methods is hard to judge.

Notes:

In our experience if a clustering implementation results in a valuable insight the effort may be deemed successful. Some of the issues in relation to clustering are:

- ***Scalability***

Most algorithms work well on small data sets. Clustering high volume databases with millions of records requires scalable methods

- ***Clusters with arbitrary shape***

Most algorithms use the Euclidian metric which are good for clusters with spherical shapes and similar size.

It is important to develop clusters for arbitrary shapes.

- ***High dimensionality***

A database contains several *attributes* or *dimensions*. Many algorithms are good at handling low-dimensional data. Higher dimensional data with data sparsity is a definite problem in clustering.

Notes:

- ***Noisy data***

Data is often noisy. Real-world databases contain outliers, missing data, erroneous data. Methods ought to deal with these conditions in the data. For further advanced research and study, the reader is referred to the list of references.