Cluster Analysis

The term ***cluster analysis*** encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories.

Cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist.

1 Cluster analysis

The term cluster analysis (first used by Tryon, 1939) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies.

In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist.

We deal with clustering in almost every aspect of daily life. For example, a group of diners sharing the same table in a restaurant may be regarded as a cluster of people. In food stores items of similar nature, such as different types of meat or vegetables are displayed in the same or nearby locations. There is a countless number of examples in which clustering plays an important role. For instance, biologists have to organize the different species of animals before a meaningful description of the differences between animals is possible. According to the modern system employed in biology, man belongs to the primates, the mammals, the amniotes, the vertebrates, and the animals. Note how in this classification, the higher the level of aggregation the less similar are the members in the respective class. Man has more in common with all other primates (e.g., apes) than it does with the more ”distant” members of the mammals (e.g., dogs), etc. In short, whatever the nature of your business is, sooner or later you will run into a clustering problem of one form or another.

k-means is a method of clustering optimization where the number of clusters must be known a priori. If you're not sure about the number of clusters you can run several FASTCLUS procs with different K values to determine which number might be the optimal. You can also try a hierarchical method with PROC CLUSTER which will give you the best mathematical solution using a distance metric. Keep in mind that PROC CLUSTER calculates the distance between all observations and all variables in a Dataset, so if it is very large you might not be able to run the PROC.

When we run FASTCLUS with different k values, it'll return output with different number of clusters everytime. So how do we know which of the outputs is the optimal?

That's the most difficult part of unsupervised learning techniques. You must use your business knowledge, the problem context or discuss the results with a subject matter expert to decide which solution might be the best. For example in a marketing scenario you might be restricted to a maximum of 6 clusters because they can't implement more targeted campaigns than that. So you should generate several possible solutions and discuss which is the most fitting solution for your problem.

You can tell PROC FASTCLUS how many clusters (i.e. K) by using the MAXCLUSTERS option, like this:

proc fastclus data=YourDataSet maxclusters=3 out=Clusters; *maxclusters= of clusters used; var YourFirstVariable YourSecondVariable EtcVariables; run;

I like to try a few different numbers of clusters and inspect them visually to see if there is a ”natural” number of clusters that forms. You can do this with the following code in SAS: /* Obtain the principal components for the same variables used in the cluster analysis */ proc princomp data=Clusters out=PrincipalComponents; var YourFirstVariable YourSecondVariable EtcVariables; run; /* Biplot: plot the clusters on the two principal components */ proc gplot data=PrincipalComponents; plot prin1*prin2=cluster; run;

The code above does two things. It first calculates the principal components using the ”Clusters” data set you made in the PROC FASTCLUS statement. You can think of principal components are variables that are each a unique combination of the variables you already have. The way they are created (and this gets too complex to explain easily), the first principal component will ”explain” the greatest amount of the variance in the data, the second will explain the second greatest amount of the variance, etc. And importantly, none of the components overlap, so they each explain different parts of the data. Once you have the principal components, you can plot your clusters on the first two to create a biplot. Try this and see where your clusters end up. You might have to change the default colors with a statement like the following to see the clusters a little more distinctly. If they are separated a good amount and appear to be fairly non-overlapping, then you have a good number of clusters. If they are overlapping or awkwardly mixed, try a different number of clusters. symbol1 v=plus c=blue; symbol2 v=plus c=red; symbol3 v=plus c=green;