Getting to Know Your Customers by Clustering on Product Purchase Patterns

James Shields, IDEXX Labs, Inc., Westbrook, ME

ABSTRACT
Often in a corporation customers are classified based on gut feel or simple sales rankings. However, clustering can assign the customers to groups based on their behavior and purchase patterns and take some of the guess work out. This enables of the customizing of treatments, such as promotion frequency, to optimize sales within clusters or it also can identify unexpected influences on sales strategies.

This presentation covers some of the basics of clustering using SAS® by illustrating several examples useful in business.

INTRODUCTION
Clustering is a method of grouping customers together so that the customers within a group, or cluster, are more similar to each other than to customers outside the group. That is, the difference between members of different clusters is greater than the differences between members of the same cluster.

To accomplish this there is usually a ‘distance’ measure where each customer is represented by a vector where the customer characteristics are the components of the vector. For instance, if

Customer j = [Xj, Yj],

where Xj, Yj are customer characteristics, then the distance between customer 1 and customer 2 is the square root of (X1 - X2)² + (Y1 - Y2)². This is a common definition of distance. This Euclidean distance measure is the easiest to understand. However, there are many different definitions of distance and there are also many techniques for joining ‘close’ customers together to form clusters.

For the examples in this paper I am using Proc Modeclus which is a non-parametric clustering technique using density estimates. Each point is assigned a density based on the number of points in a sphere defined around that point. The density is higher when there are more data points or the sphere is smaller. In a two dimensional example, if density is thought of as elevation, then a point is joined to a cluster if the path to that cluster is uphill and the path to that cluster is the steepest.

Non-parametric clustering methods have an advantage in that the resulting clusters are not biased towards any particular shapes. Many parametric techniques produce clusters that are roughly spherical. The ability to form irregularly shaped clusters allows the ability to detect subtle differences in the data.

GEOGRAPHIC CLUSTERING
As a simple, but useful, example let’s look at clustering on geographic density. In this case the customer characteristics are latitude and longitude. We can use a customer location to define groupings of customers that are relatively close to each other. Using latitude and longitude allows the easy visual exploration of cluster shapes formed by using different parameters or techniques. But this is left to the reader.

The map at the top of the next page shows customers in NJ and NY. Clustering breaks the up the customer location into three groups. The three groups make since because the points of highest density are around population centers and the clusters show this.
The clusters were defined by the following:

```
Proc Modeclus Data=input_lat_long Method=1 K=20 out=in.outclus;
where STATECODE in ('PA','NJ');
var x y;
run;
```

"Method=" specifies which of the seven clustering techniques computed by Proc Modeclus are to be used. "K=" specifies that the size of the spheres for finding the density are defined so that 20 data points are included. This means that the radius is changed to include 20 points as opposed to the density being computed based on the number of points within a sphere of fixed radius.

This is a simple two dimensional cluster. If we add another variable we can illustrate how this can be used an analyzing customer data. Let’s add another variable that represents product ownership. For simplicity sake the variable is one for customers in New Jersey and 0 in Pennsylvania. The clustering now looks like this:
Since the new variable is discrete and latitude and longitude are continuous it has a strong impact on the clustering. But it illustrates a point in that it breaks up the cluster in Southern Jersey into two parts and the one in Pennsylvania spreads completely across the state. If we did not know how the 'product' variable was set up the pattern might indicate a difference in sales between the states – perhaps due to sales force allotment or different state taxes. If the 'product' variable were continuous, perhaps sales volume, the breaks may indicate areas with similar sales volume and clusters with low sales volume may indicate areas that are being neglected by the sales force or just have lower sales potential.

CLUSTERING TO FIND THE GOLD

A business may want to use clustering to differentiate between the best and the worst customers. For instance, if we have data showing order level sales and promotion for individual customers, we may want to break this information into variables that represented different 'dimensions'. In the example above the dimensions are latitude, longitude, and product.

If we use strongly correlated sales data the clusters are frequently un-interesting. For instance, if we use sales for product one and sales for product two and the correlation is strongly positive, then we are likely to see a clustering that represents high volume of sales versus low volume of sales. For this you might as well take the top 10% of customers and call it segment 1, the next 10% and call it segment 2, and so on.

However, there may be a benefit in defining variables that may or may not be correlated, but represent unique dimensions. One example would be variables such as sales, profit, likely hood to defect, and life time value. But coming up with all these variables may be difficult. So for this next example the variables sales, median dollars per order, median time between orders, time as a customer, and number of product lines purchased are used.

We are now clustering on six dimensions since there are six variables. To do this we use Proc Modeclus in this format:

```plaintext
proc Modeclus data=input_sales_data method = 1 R=1.6 std test join=0.05
    out=in.outclus;
    var sales
        median_dollars_per_order
        median_time_between_orders
        time_as_a_customer
        number_of_product_lines
        number_of_orders ;
    ID customer_id;
run;
```

There are a few more commands used with Proc Modeclus. 'ID' tells the procedure to retain the customer_id on the output file “in.outclus”. This way we can tie a customer back to which cluster it is in. Also, instead of using just K= to define the neighborhoods we are also using R= which is used to create a fixed radius sphere instead of one defined by the k-nearest neighbors. Part of the output looks like this:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Frequency</th>
<th>Boundary Frequency</th>
<th>...</th>
<th>Z</th>
<th>Approx P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8420</td>
<td>396</td>
<td>...</td>
<td>26.858</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7144</td>
<td>742</td>
<td>...</td>
<td>20.085</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3006</td>
<td>263</td>
<td>...</td>
<td>17.043</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2195</td>
<td>594</td>
<td>...</td>
<td>5.849</td>
<td>0.001</td>
</tr>
</tbody>
</table>

The 'Test' option with the Proc Modeclus specifies that significance levels be calculated to test if there are significance differences between clusters. The "Join" option then specifies that non-significant clusters, for the
specified p-value, be joined together. For this example five clusters were defined on the first pass, not shown here, with the fifth cluster not being significantly different from cluster 2. This cluster was then merge with cluster 2 and the resulting 4 clusters were then significantly different and the process was stopped.

The boundary frequency is the count of data points on the boundary with another cluster. The better defined the clusters the closer this number will be to zero. As you can see the boundary numbers not being zero does not mean the clusters are not significant.

Now we have four well defined clusters based on these variables. But what variables are making a difference? How are the clusters different? With six dimensions it is not easy to simply plot the cluster membership and visually see the differences as was done with the geographic example. There are several methods with which to find the factors that contribute the most to a cluster. One method is using classifications trees. With SAS Enterprise Miner it is easy to get the relative importance of each variable. An analyst might look at means, which are highly influenced by outliers, or perhaps medians, for each of the variables. However, I prefer the simple method of using box plots. This is a nice visual rendering of means, medians, maximums, minimums, and even the number of customers in each group.

```
Proc Boxplot data=in.outclus;
  title1  height=5 'Time as Customer';
  plot   time_as_a_customer *cluster /
    boxstyle      = schematic
    cframe        = white
    cboxes        = black
    cboxfill      = ywh
    idcolor       = salmon
    boxwidthscale = .5
    boxwidth = 5
  ;
  label cluster="Cluster";
  label time_as_a_customer ="Days Since First Purchase";
run;
```

At the top of the next page is the output from the above code. Clearly there is a difference between cluster 1 and cluster 4 with respect to time. Not shown here are the box plots for the other variables, when looking at the box plots it becomes clear that cluster 4 is formed around the best customers who bought early on and the worse customers are in cluster 1 and only recently became customers.

From a business point of view we may want to consider concentrating on retention for segment 4 and work on building further relationships and business with the other segments to move them up to the same quality as those in segment 4. On the other hand, perhaps concentrating on retention for cluster 4 is not the way to go. These are long term customers and may be less likely to defect to the competition. Perhaps cluster 1 customers are on the low end of performance because they are not loyal and retention efforts should be placed here. Whatever the reasons we now have 4 groups of customers that can be analyzed further to determine the right allocation of resources. Perhaps response models should be built for each of these groups to optimize direct mail campaigns. But one thing to consider is that clustering does not tell you who is likely to respond. It gives you a current state of the business and is not necessarily a predictor of future performance.
CLUSTERING ON TIME

Just about any set of numeric variables can be used in cluster analysis. One use for clustering is to find patterns in purchasing over time. This can be helpful in optimizing time series models or just to determine which customers like to buy at certain times. Consider product sales over time shown in the graph below:

The total number of units sold shows a seasonal pattern but no strong trend.
Cluster analysis offers a wealth of opportunities for investigating a customer database for trends over time, geographic patterns, or identifying the profitable customers from the un-profitable. Distance can be thought of in many ways and as such the possibilities are almost limitless.

However, there are a few things that should be kept in mind. Cluster analysis is not a targeting tool. Just because a group of customers is identified as the best customer does not mean that you should focus efforts on these customers. It does not predict future performance of a cluster but produces a snap shot of the current state. Other predictive analytic methods should be used to leverage the results of clustering.

Carefully consider the variables used and the desired end result. Clustering on highly correlated variables that do not define dimensions that are significantly differently may only result in common sense customer groups. The result may be high volume customers versus low volume customers. Why use clustering for this? But well defined variables can result in very informative and unexpected clusters.

Also note that Proc Modeclus produces clusters that can be oddly shaped. This is great for exploring data. But if you want to predict where new un-assigned customers fall it may be difficult to place them in the right cluster. It may be better to use another method, such as k-means, that produce groupings that are easy to place new customers into.
REFERENCES
Ravindra Khattree, Dayanand N.Naik. Multivariate Data Reduction and Discrimination with SAS® Software
SAS Institute Inc., Cary, NC, USA

SAS Institute Inc. ‘SAS Online Help and Documentation”, Version 9.1.3

CONTACT INFORMATION
For comments and questions contact the author at:
James Shields
IDEXX Labs, Inc..
1 IDEXX Lane
Westbrook, ME
Phone: 207-856-8182
Email: James-Shields@idexx.com

TRADEMARKS
SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries ® indicates USA registration.