Creating Customer Segments

Feature Transformation

The first few PCA dimensions will be those that account for the highest proportion of variation in the output. We would expect them to be composed of original features that are correlated in some way: perhaps they are combinations of items that customers tend to purchase together, such as 'fresh' and 'grocery' products. The ICA dimensions will be dimensions that are independent of one another. It seems like 'delicatessen' and 'detergents_paper' are classes that would not be strongly correlated to other classes insofar as customer purchases are concerned, so they would likely show up as ICA dimensions.

The principal component dimensions represent vectors that are linear combinations of the original feature dimensions. In fact, the analysis above shows them to be mutually orthogonal unit vectors. We can think of the six principal component vectors as elements of a principal component matrix, which because the individual vectors within it are orthonormal is orthogonal. This matrix corresponds to a set of six-dimensional orthogonal axes rotated with respect to the original six-dimensional feature axes. The principal components matrix can be applied to an input feature vector to project it onto the principal component axes. Now for our purposes it's not so interesting to take an input, rotate it around in six-dimensional space and see what it looks like when written in terms of a principal component system. But we can see how to extend the concept to execute the highly useful dimensionality reduction discussed in question 2. As opposed to a 6x6 principal component matrix that projects a 6-dimensonal-feature vector onto a 6-dimensional principal component space, we can construct a 2x6 principal component matrix that projects a 6-dimensional feature vector onto a 2-dimensional principal component space. Provided this reduced dimensionality principal component space explains a sufficient amount of the variance in the data (which by the analysis above, we know ours does) we can substantially reduce the computational workload a clustering algorithm would be asked to perform without sacrificing much in terms of valuable information.

The first principal component dimension has a coefficient of -0.98 for the 'fresh' feature, -0.15 for the 'frozen' feature and so on. It's pretty obvious that the first principal component is predominantly the 'fresh' feature vector rotated a bit away from the original. Despite the answer given to question 1 above, it seems like the 'fresh' feature has become a principle component almost unto itself. And now that we see this, we see the error in our ways that led to the erroneous answer to question 1. There is no reason that the most principal component need elucidate a strong covariance relationship between features. According to the README doc, the 'fresh' feature has the largest variance of any feature, and we attribute this, at least in part, to scale: the shops spend 50% more on average on 'fresh' items than on those from any other category. There is no need for 'fresh' to correlate to any other feature to have a great influence on total variance. There just needs to be a lot of variation from shop to shop on how much they spend on 'fresh' items. We can use this finding to suggest our wholesale distributor client increase marketing on 'fresh'

products to shops who don't buy much of it, as there seems to be a great but uneven demand for 'fresh' products across the client's customer base.
The second dimension shows what appears to be the type of correlation we initially expected to see, between two or even three of the original features: 'grocery' with coefficient 0.76, 'milk' with coefficient 0.52 and perhaps 'detergents_paper' with coefficient 0.37. This component is much more a composite of multiple features than the first principal component. This is an interesting finding, as it tells us that stores that buy more 'grocery' items tend also to buy more 'milk' items and to some degree more 'detergents_paper' items as well (and they do correlate positively: all coefficients are the same sign: positive). We can use this discovery to encourage our wholesale distributor client to try to increase sales by marketing these items in a synergistic fashion, and to reduce costs by bundling/delivering them together.

Clustering

The graph above shows the customer data plotted in terms of the first principal component along the x-axis and the second principal component along the y-axis. The various colors represent the optimal clusterings.

Recall from question 2 that the first principal component is predominantly the negative of the 'fresh' category, and that the second principal component is an amalgamation of 'grocery', 'milk' and a bit of 'detergents_paper' (hereafter known as 'GMDP'), all positive. Since the x-axis shows negative values, the larger magnitudes are to the left.

The number of clusters was determined by an ad-hoc grid search using Bayesian Information Criteria (BIC) above, Details of the clusters, including sample size and average purchases by product group were provided earlier. Here we discuss them in terms of the plot immediately above, and introduce a bit of conjecture as to what type of businesses these customers might be.

Cluster 0: the small customers that buy mostly 'GMDP' and don't buy much 'Fresh' or other products, shown in the graph as a vertical sky blue oval in the lower right. These might be convenience stores and corner markets.

Cluster 1: the small customers that buy mostly 'Fresh' and don't buy much 'GMDP' or other products, shown in the graph as a long and narrow bluish green horizontal oval in the lower right. These might be delis and single-outlet fast food shops.

Cluster 2: the mid-sized customers that buy mostly 'Fresh' but also buy a good amount of 'GMDP' and other products, shown in the graph as the large greenish spearhead shape. These might be larger restaurants.

Cluster 3: the customers who tend to buy lots of both 'fresh' and 'GMDP', whose area takes up most of the diagram and is shown in red. These customers might be big-box retailers who sell a wide variety of products in bulk.

Cluster 4: the large customers who tend to buy mostly 'fresh' products, shown along the bottom toward the left in orange. These might be cafeterias belonging to large corporations, schools or the government (military bases, prisons, etc.).

Cluster 5: a singe customer that buys a huge amount of 'Fresh' products. It does not have a corresponding area on the color map, but is represented by the white cross at the extreme left. It might a regional distribution center for a national fast food chain.

Cluster 6: the small customers that buy mostly 'Fresh' as well as a decent amount of 'GMDP' or other products, shown in the graph as the lavender spearhead shape in the lower right. These might be delis and single-outlet fast food shops, similar to Cluster 1 customers but perhaps serving different neighborhoods or customer bases.

Cluster 7: the mid sized customers that tend to buy predominantly 'GMDP' products, shown as the occluded brown semi-egg at right. These might be mid sized single-outlet grocery stores.

Conclusions

The technique that gives proper insight to the data it seems that the Gaussian Mixture Model helped a lot to make sense out of the data. It tended to identify customer groups that lent themselves to easy (and hopefully correct) interpretation. We suspect it has done a better job at elucidating these details then K Means would, as we would not expect K Means to identify the long and narrow clusters that are nicely picked out by the Gaussian Mixture Model. Of course, all this was enabled by PCA which did an excellent job of collapsing six dimensions of data down to a convenient but still highly informative two. Now that we have demonstrated a method of categorizing our client's customers based on what they purchase and how much of it, we can propose modifications and improvements to our client's delivery and marketing strategies that are tailored to a customer group. For example, while a low-cost bulk nightly delivery schedule was not acceptable to our customer's smaller clients, we could propose experiments to identify reduced cost delivery methods that would be acceptable to these customers, perhaps a bulk delivery system in the early business hours. We could also run some marketing experiments, like offering discounts on 'milk' products to customers who buy a lot of 'grocery' products to leverage product synergies and increase overall sales.