Project 3: Creating Customer Segments

Robert Crowe

Robert@ourwebhome.com

15 August 2016

Question 1 - Choose and characterize 3 examples

Consider the total purchase cost of each product category and the statistical description of the dataset above for your sample customers.

What kind of establishment (customer) could each of the three samples you've chosen represent?

Answer 1

	Row	Fresh	Milk	Grocery	Frozen	Detergents Paper	Delicatessen
0	421	17063	4847	9053	1031	3415	1784
1	162	15177	2024	3810	2665	232	610
2	182	694	8533	10518	443	6907	156

0 (421) could be a large cafe that offers menu choices with a lot of vegetables as well as some sandwiches. They are one of the larger buyers of fresh food, just above the 3rd quartile. They buy just above the median in most other categories - milk, grocery, and Detergents/Paper. They buy below the median in frozen food, and just below the 3rd quartile in deli. Apparently they use a lot of fresh food and deli.

1 (162) could be a restaurant that is mainly focused on sit-down dinners. They are a fairly balanced buyer, buying just above the median but below the 3rd quartile in both fresh and frozen, and below the median but above the 1st quartile in milk, grocery, and deli. They are a smaller buyer of Detergents/Paper, below the 1st quartile. They seem to focus mainly on food items, and need only a smaller amount of cleaning and paper goods.

2 (182) could be a small market that offers some packaged food to-go. They buy very little fresh food, well below the 1st quartile, and little frozen and deli, also below the 1st quartile. They buy a lot of milk and detergents/paper, well above the 3rd quartile, and their purchases of grocery are above the median and just below the 3rd quartile. A small corner market often sells milk, packaged food, detergents and

paper goods, but many customers will go to a larger market for things like fresh food, frozen, and deli, where they will find a larger selection.

Question 2 - Feature Relevance

Which feature did you attempt to predict? What was the reported prediction score? Is this feature necessary for identifying customers' spending habits?

Answer 2

I thought that it was worthwhile to check all 6 features, because it was easy to do and there are only 6.

Fresh Score = -0.381145015025
Milk Score = 0.171822424067
Grocery Score = 0.689154638372
Frozen Score = -0.246524772666
Detergents_Paper Score = 0.398788736906
Delicatessen Score = -2.2547115372

The grocery score is the most highly correlated with the other features of the data, with an R2 score of 0.68, and is the most easily predicted from the other scores. That means that the additional information that we get from the grocery score alone is the least of all the features, since we can get the majority of it from the information in the other features. A customer's purchases of groceries doesn't tell us much about their overall spending habits. Detergents/paper is also fairly highly correlated with an R2 score of almost 0.4.

The deli score is the least correlated with the other features, with an R2 score of -2.25, making it the most difficult to predict from the other features. That means that the additional information that we get from the deli score is the highest of all the features, since we get very little of it from the information in the other features. It is the most independent feature in the data, and could easily resemble one of the independent components if we were doing independent component analysis. A customer's purchases of deli items tells us a lot about their overall spending habits.

The most necessary feature for identifying a customer's spending habits is the deli feature, since that contains the highest amount of unique information of any of the features.

Question 3 - Correlation Between Pairs of Features

Are there any pairs of features which exhibit some degree of correlation? Does this confirm or deny your suspicions about the relevance of the feature you attempted to predict? How is the data for those features distributed?

Answer 3

Grocery and Detergents_Paper are fairly highly correlated with each other. That probably makes sense, since Detergents_Paper is often offered and purchased retail at the same type of store – a "grocery store" – and although this is a wholesale business it probably still makes sense. This confirms the result above, that grocery and Detergents_Paper are poor differentiators. The data in the scatter plot for these two features are tightly distributed around a linear regression line. Milk is also fairly highly correlated with both grocery and Detergents_Paper.

The Kernel Density Estimation (KDE) plots along the diagonal show that the distributions for all of these features are strongly right-skewed. That means that the most frequent purchases (the mode) are below the average purchase (mean), and the middle of the distribution (median). That shows that the majority of customers purchase some minimum amount of all of these items, but some purchase significantly more, drawing the right tail of the distribution out to the right. This is more true for some features than others. Fresh has a fairly long right tail, meaning that the mean and median are more significantly above the mode, meaning that there are relatively more customers who are making larger purchases of these items. Deli on the other hand has a very short right tail, meaning that there are only a few customers who purchase significantly more deli.

Question 4 - Outliers

Are there any data points considered outliers for more than one feature based on the definition above? Should these data points be removed from the dataset? If any data points were added to the outliers list to be removed, explain why.

Answer 4

There were a relatively small number of data points that were outliers in two or more dimensions (rows 65, 66, 75, 128, and 154, only 5 out of 440). Only one (154) is an outlier in more than two dimensions, but being an outlier in 3 dimensions means that it is an outlier in half of the dimensions, since there are only 6 dimensions altogether. With so few dimensions, and so few outliers, I felt that it was worth the relatively small loss of information to gain the benefit of cleaning up the distribution, so I removed them.

Question 5 – Explained Variance

How much variance in the data is explained in total by the first and second principal component? What about the first four principal components? Using the visualization provided above, discuss what the first four dimensions best represent in terms of customer spending.

Hint: A positive increase in a specific dimension corresponds with an increase of the positive-weighted features and a decrease of the negative-weighted features. The rate of increase or decrease is based on the individual feature weights.

Answer 5

The 1st and 2nd principal components explain 0.447 + 0.2746 = 0.7216 or roughly 72% of the total variance in the data.

The first 4 principal components explain 0.447 + 0.2746 + 0.1123 + 0.1004 = 0.9343 or roughly 93% of the total variance in the data.

The first dimension is dominated by Detergents_Paper, with strong contributions from Milk and Grocery. I'm not sure exactly what Grocery includes, but since it excludes Fresh, Frozen, and Milk it would seem to be things like canned and packaged food items, Bread, drinks other than Milk, and probably a wide variety of other items. So basically the first dimension seems to be everything except fresh and frozen food, and only small amount of deli.

The second dimension is almost the opposite of the first. It is dominated by fresh food, frozen food, and deli. It seems to be food, but excluding packaged foods like cookies or chips, and most drinks. The contribution from Milk could perhaps be milk used in cooking.

The third dimension is dominated by fresh food, with a contribution from Detergents_Paper, and a strong aversion to deli. It could easily be fresh food used for cooking dishes which are then served on paper plates with paper napkins, and a bit of soap for cleaning up. It could perhaps be delis preparing their own deli items, and not interested in buying deli items already prepared.

The fourth dimension is dominated by frozen food almost entirely, with aversions to deli and fresh food. This could perhaps be frozen food for resale as frozen food.

Question 6 – Choose a Clustering Algorithm

What are the advantages to using a K-Means clustering algorithm? What are the advantages to using a Gaussian Mixture Model clustering algorithm? Given your observations about the wholesale customer data so far, which of the two algorithms will you use and why?

Answer 6

K-Means scales well as the number of samples grows, and especially if Mini Batch K-Means is used to sample from the whole dataset it scales reasonably well as the number of clusters grows. Although theoretically K-Means is an $O(K^n)$ algorithm, in practice since the error decreases with each iteration the actual complexity is generally much lower than that.

In contrast, Gaussian Mixture Models do not scale well. It also has a problem that "when one has insufficiently many points per mixture, estimating the covariance matrices becomes difficult, and the algorithm is known to diverge and find solutions with infinite likelihood unless one regularizes the covariances artificially" while K-means is known to always converge, although it can converge to a local minimum so it is important to run it more than once and randomly initialize the centroids when using K-Means.

KMeans can be seen as a special case of Gaussian Mixture Model with equal covariance per component. However, the principal components of the wholesale customer data (before dimensionality reduction) do not have equal covariance, so KMeans would probably not be the best choice. A Gaussian Mixture Model incorporates information about the covariance structure of the data, as well as the centers of the latent Gaussians. Therefore, a Gaussian Mixture Model is a better choice for finding the clusters in the wholesale customer data.

Question 7 – Silhouette Score

Report the silhouette score for several cluster numbers you tried. Of these, which number of clusters has the best silhouette score?

Answer 7

Silhouette score with 2 components and spherical covariance is: 0.333965327585
Silhouette score with 3 components and spherical covariance is: 0.390318537174
Silhouette score with 4 components and spherical covariance is: 0.356229447193
Silhouette score with 5 components and spherical covariance is: 0.367451676442
Silhouette score with 6 components and spherical covariance is: 0.284614475553
Silhouette score with 7 components and spherical covariance is: 0.347588897941
Silhouette score with 8 components and spherical covariance is: 0.31547580718
Silhouette score with 9 components and spherical covariance is: 0.326612987774
Silhouette score with 2 components and tied covariance is: 0.377861216813
Silhouette score with 3 components and tied covariance is: 0.358622190354
Silhouette score with 4 components and tied covariance is: 0.343680865088
Silhouette score with 5 components and tied covariance is: 0.344842173563
Silhouette score with 6 components and tied covariance is: 0.300483422134
Silhouette score with 7 components and tied covariance is: 0.335798564189
Silhouette score with 8 components and tied covariance is: 0.302785672861
Silhouette score with 9 components and tied covariance is: 0.327124728156
Silhouette score with 2 components and diag covariance is: 0.371936042091
Silhouette score with 3 components and diag covariance is: 0.393990121348
Silhouette score with 4 components and diag covariance is: 0.326967014792
Silhouette score with 5 components and diag covariance is: 0.305183130734
Silhouette score with 6 components and diag covariance is: 0.269289784587
Silhouette score with 7 components and diag covariance is: 0.275286576257
Silhouette score with 8 components and diag covariance is: 0.235738970487
Silhouette score with 9 components and diag covariance is: 0.307387260296
Silhouette score with 2 components and full covariance is: 0.373468456625
Silhouette score with 3 components and full covariance is: 0.359248611801
Silhouette score with 4 components and full covariance is: 0.284581439039
Silhouette score with 5 components and full covariance is: 0.326372301666

```
Silhouette score with 6 components and full covariance is: 0.287249889333
Silhouette score with 7 components and full covariance is: 0.196668252407
Silhouette score with 8 components and full covariance is: 0.132546128035
Silhouette score with 9 components and full covariance is: 0.260146307447
```

Using the silhouette score to select the best model, the best number of components is 3, which gives a score of 0.3939. However, a case could be made for using the Bayesian Information Criterion (BIC) to select the best model, which would result in a silhouette score of 0.3674, nearly as good as what we get by maximizing the silhouette score, but with a best number of components of 5 rather than 3. The higher number of customer segments when using BIC may be more useful for the wholesale distributor. But since the assignment focuses on using the silhouette score, I'll use the silhouette score.

Question 8 – Cluster Archetypes

Consider the total purchase cost of each product category for the representative data points above, and reference the statistical description of the dataset at the beginning of this project. What set of establishments could each of the customer segments represent?

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Segment 0	8,519	6,130	8,385	1,800	2,217	1,399	28,450
Segment 1	8,885	1,801	2,316	2,060	276	678	16,016
Segment 2	2,172	4,475	7,178	606	2,107	487	17,025

Hint: A customer who is assigned to 'Cluster X' should best identify with the establishments represented by the feature set of 'Segment X'.

Answer 8

- Segment 0 looks to me like a larger produce market, who is buying from this wholesaler. They're reselling some of a typical assortment of food and grocery items, but their main focus is clearly fresh food, where they are just above the median. The other categories besides fresh are down around the 1st quartile, so they are selling some but they aren't their main focus.

- Segment 1 looks like a large grocery store. They are also just above the median for fresh food, but close to the 3rd quartile for both milk and grocery. They also sell just above the median in frozen, detergents/paper, and deli.
- Segment 2 looks to me like a small corner market, who is selling a typical assortment of food and grocery items with a higher mix of packaged foods (grocery) than the grocery stores in segment 0. Their total purchases are smaller than the other two segments, and their main focus is grocery, where they are well above the median.

Question 9 – Segments for Sample Points

For each sample point, which customer segment from **Question 8** best represents it? Are the predictions for each sample point consistent with this?

Answer 9

Sample point 0 predicted to be in Cluster 0 Sample point 1 predicted to be in Cluster 1 Sample point 2 predicted to be in Cluster 2

Well, it looks like I got lucky and chose a sample point from each of the segments.

My prediction for sample point 2 was pretty close, but for sample point 0 I think the larger produce market hypothesis is probably better than the "large cafe that offers menu choices with a lot of vegetables as well as some sandwiches", and for sample point 1 I think the large grocery store hypothesis is probably better than a restaurant. I just can't imagine a restaurant ordering that much in groceries, although I suppose that they could be using a lot of packaged foods.

To really pin this down I would want to go back to the original customer list and try to study and characterize the customers in each segment. Trying to rely solely on the purchase data probably isn't the best way to understand a customer segment.

Question 10 – A/B Testing

Companies often run A/B tests when making small changes to their products or services to determine whether that change affects its customers positively or negatively. The wholesale distributor wants to consider changing its delivery service from 5 days a week to 3 days a week, but will only do so if it affects their customers positively. How would you use the customer segments you found above to perform an A/B Test for this change?

Hint: Can we assume the change affects all customers equally? How can we determine which group of customers it affects the most?

Answer 10

In reality reducing the availability of delivery service is unlikely to affect customers positively unless there is some benefit, like a reduced cost of delivery. The most realistic goal is that reducing delivery options would not affect customers too negatively. But for the sake of argument let's assume that it's possible that having fewer options for delivery could somehow affect customers positively.

Also in reality the wholesaler would have information about when customers are currently getting deliveries, and so we could focus testing on those customers who would see the biggest changes in their delivery schedules. But since that data isn't part of the project, let's assume that it's not available.

To use the segments to do A/B testing we should pick a statistically significant sample from each of the segments and try reducing deliveries to only 3 days for those samples. After a period of at least 1 week, and probably more like 2-3 weeks, we should measure customer satisfaction among those sampled and compare to customer satisfaction among the control group, which are the customers within each segment who are still receiving deliveries all 5 days. We should also look at purchase behavior among the two groups to see if purchases among the test group have increased, decreased, or remained steady, and how that compares to trends within the control group.

The main point here is to compare customers within each segment, rather than comparing customers between segments.

Question 11 – Engineered Features and Classifying New Customers

Additional structure is derived from originally unlabeled data when using clustering techniques. Since each customer has a segment it best identifies with (depending on the clustering algorithm applied), we can consider 'customer segment' as an engineered feature for the data. Assume the wholesale distributor recently acquired ten new customers and has made estimates for each customer's annual spending of the six product categories. Knowing these estimates, the wholesale distributor wants to classify each new customer to one of the customer segments to determine the most appropriate delivery service.

Describe a supervised learning strategy you could use to make classification predictions for the ten new customers.

Hint: What other input feature could the supervised learner use besides the six product features to help make a prediction?

Answer 11

Using the features of the current customers as the training set, and using the segments as the labels, we can train a classification model. Then using the estimates of the spending for the new customers as features, we can use the model to predict the segment that each of the new customers falls into.

Question 12 – Compare Segments to Channels

How well does the clustering algorithm and number of clusters you've chosen compare to this underlying distribution of Hotel/Restaurant/Cafe customers to Retailer customers? Are there customer segments that would be classified as purely 'Retailers' or 'Hotels/Restaurants/Cafes' by this distribution? Would you consider these classifications as consistent with your previous definition of the customer segments?

Answer 12

None of the segments are purely Hotel/Restaurant/Cafe customers or Retailer customers, although cluster 1 is mostly Hotel/Restaurant/Café, and cluster 2 is mostly Retailers. Cluster 0 is a mix of both, but exhibits different purchase behavior that either, which probably indicates other underlying factors that are important to their business.

The division between the purchase behavior of Hotel/Restaurant/Cafe customers and Retailer customers fairly closely mirrors the division between cluster 1 and cluster 2, so that tends to validate the clustering results. The addition of the third cluster probably reveals underlying factors that are not reflected in the division between Hotel/Restaurant/Cafe customers and Retailer customers.

While the classification based solely on channel - Hotel/Restaurant/Cafe customers or Retailer customers – is somewhat consistent with the results of clustering, I think the clustering reveals better information for understanding the customers. That could easily have been the problem with the initial trial, which may have considered customers only on the basis of channel.