Creating Customer Segmentations with Unsupervised Learning Techniques

George Liu

Summary

In this project, we aim to use unsupervised learning techniques to help our client, a wholesale grocery distributor, to identify different customer segments in order to better understand customer behaviors and devise corresponding marketing strategy and operational plan to better meet customers' needs. Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Gaussian Mixture Model clustering (GMM) techniques are explored and final customer segmentations are recommended.

Component analysis

1. Potential PCA/ICA Components

   There may be two principal components. The first principal component could be average order size in terms of monetary value, since this is the major differentiating factor among all customers. The second principal component might be something less significant yet still important to identify different customers. It could be order frequency or product mix (as a ratio) etc.

   For ICA, since the algorithm further breaks the original features down into even more fundamental elements, the dimensions could be product SKU's or something similar that is lower level than product categories (i.e. fresh/milk/grocery etc.), they might be sub-categories or something similar.

2. PCA

   The following values represents the proportions explained by each principal component:

   [0.45961362 0.40517227 0.07003008 0.04402344 0.01502212 0.00613848]

   The variance drops very fast. The first two components represent roughly 87% of the variance. The third to the sixth represent about 7%, 4%, 2% and 1% respectively. So it does appear that two components represent the majority of the variance in the data, which is in line with my initial hypothesis. Therefore, I would choose two dimensions for my analysis.

Based on the result given by pca.components_, we can see that the first component is a linear combination of mainly these variables (with weights shown in brackets):

- Fresh (-0.98), Frozen (-0.15), Milk (-0.12)

Given the above combination, it seems this component represents “Horeca” type channel spending, such as hotel, restaurant and café.

The second principal component, similarly, is composed of:

- Grocery (0.76), Milk (0.52), Detergents_Paper (0.37), Fresh (-0.11)

This combination looks more like retail channel type spending. Furthermore, it is important to note that we see Fresh in both components, whereas Fresh shares the same sign with Frozen and Milk in PC-1, but have a different sign than the rest in PC-2. So, when considered together, it appears that the two principal components represent a characteristic of channel type, or “freshness”, i.e. how much fresh products are purchased, with PC-1 pointing to retail who buy mainly fresh products, and PC-2 pointing to “Horeca” channel – hotel, restaurant and cafés who buy more grocery, milk and detergents/paper products.

This transformation is useful, as we can now visualize the data in a 2-dimension space. Furthermore, with the insight of channel type in mind, we can have a rough understanding of this wholesale grocery distributor’s customer composition and behavior. For example, since the first principal component represents Fresh, Frozen and Milk (FFM), we can say that the customers mainly differ in terms of their FFM spending pattern. Secondly, the customers differ in regards to their GMDF (Grocery, Milk, Detergents_Paper and Fresh) spending pattern.

3. ICA

Here is the ICA result:

The independent components are:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
IC-1	0.003792	-0.016976	-0.114580	0.007090	0.134342	0.016148
IC-2	-0.001940	-0.072583	0.055137	0.001766	-0.015770	0.017065
IC-3	0.010930	0.001035	-0.007349	-0.054049	0.002645	0.016767
IC-4	-0.002663	0.013924	-0.060813	-0.002006	0.003626	0.004070
IC-5	0.050227	-0.006331	-0.005911	-0.003286	0.009799	-0.002942
IC-6	0.004882	0.001618	0.005705	0.002532	-0.002430	-0.050965

The first independent component (IC) is mainly concerned with Detergents_Paper and Grocery and they do have a negative relationship, i.e. when one spending is high, the other one tends to be lower. The second IC is mainly Milk and Grocery (negative relationship). The third IC is composed mainly of Frozen, Delicatessen and Fresh. The fourth IC is made up of Grocery and Milk. The fifth is mainly fresh. The last IC has major dependence on Delicatessen.

To sum up, the above data appears to support my initial hypothesis of the concept of sub-categories, since each IC is made up of one major original categories with some adjustment:

- IC-1 = Detergents_Paper - Grocery
- IC-2 = Grocery - Milk
- IC-3 = Frozen - Delicatessen - Fresh
- IC-4 = Milk - Grocery
- IC-5 = Fresh
- IC-6 = Delicatessen

The above information is really insightful for 2 reasons. One, when we know the relationship between two categories, we can potentially use that as part of our marketing efforts to increase sales. Two, the above equations reveal the independent components behind the scene for us. Originally, we think the categories such as Fresh, Milk etc. best divide up the products for us. However, ICA tells us some optimization is needed. For example, maybe some grocery products are wrongly categorized into Detergents_Paper, or perhaps some Milk products are put under the Grocery category.

Clustering

4. Clustering Methods and Model Selection

The below table provides the advantages and disadvantages of K-means and GMM clustering:

	Advantages	**Disadvantages**
K-means	Very fast (one of the fastest clustering algorithms available); Scales well to large number of samples; Will always converge given enough time	Falls in local minima (thus useful to restart several times); Responds poorly to elongated clusters, or manifolds with irregular shapes; Unreliable with high-dimensional data (Curse of Dimensionality); Hard assignment at each iteration; Needs to specify number of clusters; Linear decision boundary
	Reference: K-means on Sklearn, The Challenges of Clustering High Dimensional Data, Difference between K-means and GMM	
GMM	Fastest algorithm for learning mixture models; Does not bias means or cluster sizes, more general/less assumptions; Soft assignment of each point based on probability; Number of components can be determined using BIC or Dirichlet process; Non-linear decision boundary for better flexibility	Not scalable; Can fail with high dimensionality; Need to decide number of components; May not converge
	Reference: GMM on Sklearn, GMM Classifier, Why is the decision boundary for K-means clustering linear?	

Here, since our clusters can have irregular shapes, and in order to have a more precise clustering, we will choose GMM to proceed with model building.

To decide the best model, a model selection process that uses Bayesian Information Criterion is applied. The model with 8 components and covariance type of “diag” gives the highest BIC score. Therefore, the number of clusters should be 8.

5. Cluster Centroids

Based on the clusters made, we also find all the corresponding centroids. These centroids are the respective “central points” within the clusters. Therefore, they represent the typical customer within each segment.

Since these centroids are data points in the PCA-reduced 2D space, in order to better interpret these customer profiles, we need to put the centroids back into the original space. This is done with the inverse_transform method of PCA. Results are provided below (the numbers are the data points’ percentile ranks in the original data set columns):

	Fresh	Milk	Grocery	Frozen	Detergents_Paper
Segment A	65.227273	27.954545	23.181818	71.363636	8.181818
Segment B	17.727273	76.590909	78.409091	50.227273	80.909091
Segment C	89.090909	98.636364	98.863636	83.181818	98.409091
Segment D	95.454545	58.636364	38.409091	89.090909	0.000000
Segment E	46.590909	91.590909	90.227273	62.727273	91.590909
Segment F	34.772727	35.000000	42.045455	57.500000	55.227273
Segment G	76.590909	66.590909	64.772727	77.500000	66.590909
Segment H	100.000000	97.045455	93.181818	99.318182	83.409091

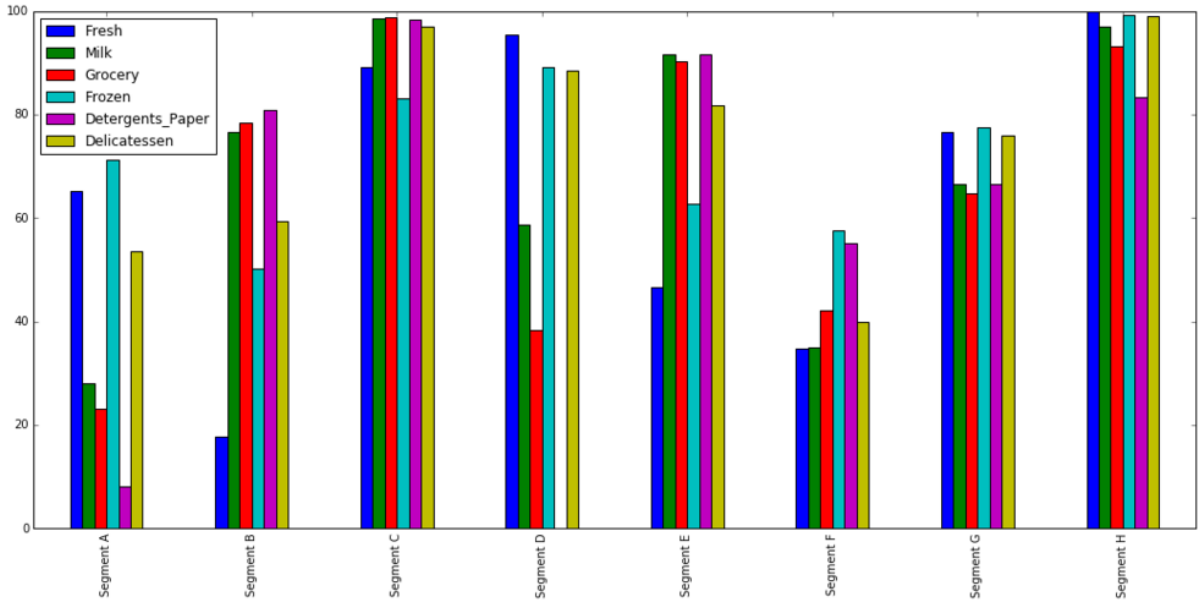
	Delicatessen
Segment A	53.636364
Segment B	59.318182
Segment C	97.045455
Segment D	88.636364
Segment E	81.818182
Segment F	40.000000
Segment G	75.909091
Segment H	99.090909

We can see the following segment profiles emerging from the clusters:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper
Segment A	high	low	low	high	very low
Segment B	low	high	high	medium	high
Segment C	high	very high	very high	high	very high
Segment D	very high	medium	low	high	very low
Segment E	medium	very high	very high	high	very high
Segment F	low	low	medium	medium	medium
Segment G	high	high	high	high	high
Segment H	very high	very high	very high	very high	high

	Delicatessen
Segment A	medium
Segment B	medium
Segment C	very high
Segment D	high
Segment E	high
Segment F	low
Segment G	high
Segment H	very high

And here is the visulization of the finding:

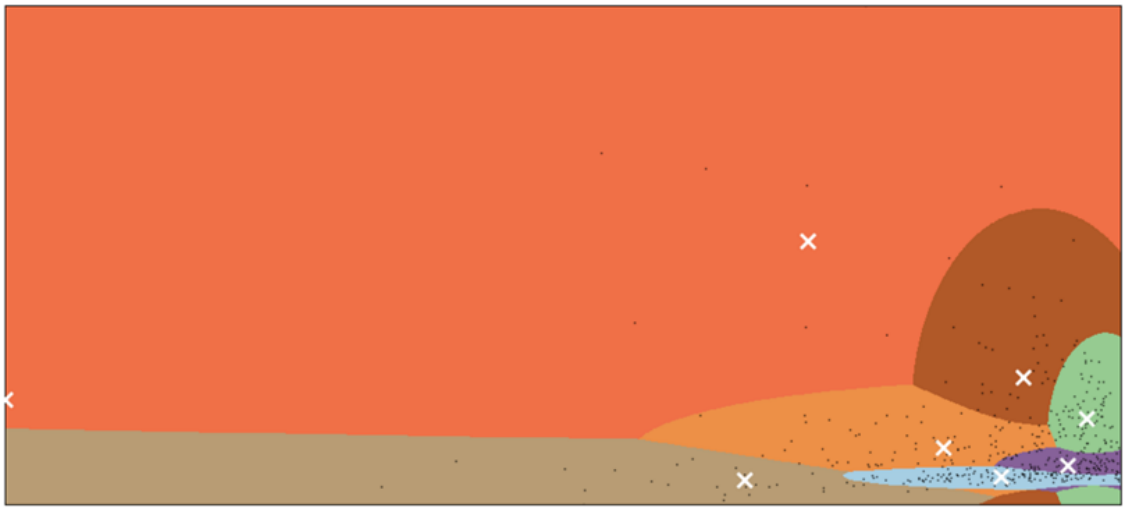


Based on this, we make some comparison with real life situations regarding the segments:

- Segment A: likely Horeca channel, appears to be restaurants
- Segment B: likely Retail channel, specific type unclear
- Segment C: likely Retail channel, appears to be supermarkets
- Segment D: likely Horeca channel, appears to be cafés
- Segment E: likely Retail channel, specific type unclear
- Segment F: likely Horeca channel, appears to be hotels
- Segment G: likely Retail channel, appears to be convenience stores
- Segment H: likely Retail channel, appears to be grocery stores

6. Visualization

We now visualize the cluster centroids in the PCA-reduced 2D space as below:



The respective cluster centroids are marked with white X in the graph. We can see that the clusters produced by the GMM algorithm is much more sophisticated than that of K-means clustering, as the boundary can be curves instead of straight lines.

Some of the clusters are not very clear. There are several things we can do to improve:

- Remove the centroid marker to give more room
- Optimize colors used for higher visibility
- Normalize or log transform the data for more "spread out"

Conclusions

We used PCA, ICA and PCA + GMM clustering in this project. Out of these techniques, the combination of PCA and GMM clustering appears to provide the most insights. Our goal is to find out the underlying customer segmentations, i.e. the different groups of customers who share similar purchasing behaviors. PCA gives us a good understanding of the major driving force behind the scenes, and also allows us to reduce the dimensionality of the data so that we can visualize it easily. ICA provides interesting findings about the potential new ways of product category labeling. However, it is the combination of PCA and GMM clustering techniques that best serves our purpose to identify customer segments. This is due to the fact that clustering techniques naturally groups customers by similarity, and PCA supplies dimension-reduced data for GMM to work optimally.

The findings can be very useful for our grocery client. Different customer segments exhibit different behaviors and have different needs. For future marketing or operational experiments, instead of picking random candidates to test out, our client can instead conduct pilot tests with different customer segments, and only roll out when a certain plan is proved to work with a certain customer segment.

Specifically, if our client is considering making changes to the delivery method, we can advise the client to run A/B tests based on the identified customer segments, i.e. instead of testing out the new method on the whole customer base, our client can pick specific segments to test with (based on customer needs analysis). For a specific segment, a test sample can be picked and further randomly divided into control and experiment groups. For the experiment group, we'll provide the new delivery method while for the control group, we simply do nothing, i.e. to keep the existing delivery method. Finally, we can use satisfaction score obtained using customer survey as evaluation metrics to measure the effect of the test.

If the new delivery method generates statistical and practical significant preferable difference, the client can consider rolling out the change to the full corresponding segment and stay unchanged otherwise. This way, we can have confidence that new business initiatives won't have negative impact on various customer segments thanks to more granular understanding of customer differences and segmentations.

The other way that we can take advantage of the findings is to use supervised learning techniques to classify new customers into corresponding customer segments. This will allow us

to predict customer needs based on their segments behaviors and will provide opportunity to implement customized and targeted marketing and operational tactics to ensure customer loyalty while balancing business needs and profitability.

Reference:

1. What is Independent Component Analysis?
2. Independent component analysis
3. Principal Component Analysis
4. Eigenvectors and Eigenvalues
5. Principal Component Analysis and Regression in Python
6. Recovering features names of explained_variance_ration in PCA with sklearn
7. How to use scikit-learn PCA for features reduction and know which features are discarded
8. Making sense of independent component analysis
9. The Challenges of Clustering High Dimensional Data
10. GMM Classifier
11. Gaussian Mixture Model Selection
12. Wholesale Customers Data Set
13. Why is the decision boundary for K-means clustering linear?
14. Clustering