UDACITY

MACHINE LEARNING – NANODEGREE

CUSTOMER SEGMENTS

BY: TODD FARR

1 Component Analysis

1.1 Reflection on PCA/ICA

PCA or Principle Component Analysis is a technique that often allows for the dimensionality reduction of feature spaces by identifying directions that represent the highest orthogonal variance of a dataset. By identifying these high variance directions, conclusions can be made about which features contain the most "information" and which features are similar and can be combined into composite features. For this particular dataset these are features (item groups) that would often be purchased by the same types of customers like 'groceries' and 'fresh.'

ICA or Independent Component Analysis on the other-hand identifies directions or "driving sources" in the data that are statistically independent from one another. These independent components in regards to this dataset would represent hidden variables in the form of eigenvectors that we could potentially associate to different buyer types.

Reflecting a little more on these two types of analysis, they are often best understood when their interactions with a given dataset are compared visually. A 2D simple, yet effective, representation of these two algorithms and the directions they uncover can be seen in Figure 1. From this visualization it becomes a little more clear that ICA is identifying independent directions in the data while PCA is finding a direction and the associated orthogonal the best captures the highest variance in the dataset.

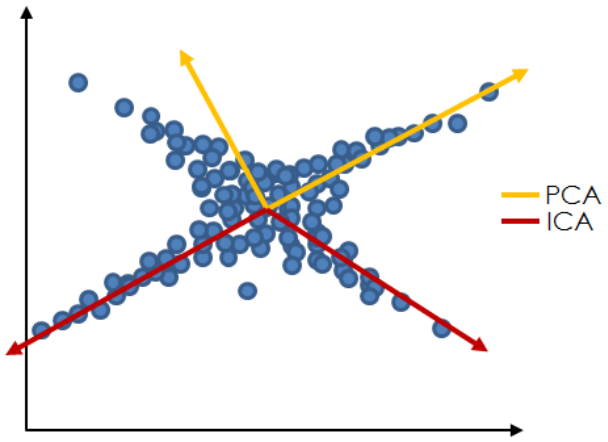

Figure 1: PCA vs ICA components visualized on an example dataset.

1.2 What proportion of variance is explained by each PCA dimension?

The eigenvalues for each PCA dimension directly reflect the amount of variance that can be explained by its eigenvector. Since the dimensions represent the entire dataset then coincidentally the summation of all eigenvectors can explain the total variance of the dataset.

1.3 PCA Dimensions

The first few components of a PCA are those features that have the highest explained variance. These components or composite features are the best at generalizing the features contained in the data while minimizing the associated information loss caused by compressing it. The most common use for this analysis technique is a data dimensionality reduction in complex feature spaces to alleviate the associated complications that arise from trying to process these datasets. In this particular instance running a PCA on the dataset reveals that approximately 86% of the variance can be explained by 2 resultant features. This results in minimal model differences for a feature space reduction from 6 to 2. This can be further visualized by plotting the individual feature PCA vectors in a two-dimensional space created from the first and second driving PCA components. It's likely the composite features are combinations of the features that point in the same directions (Figure 2).

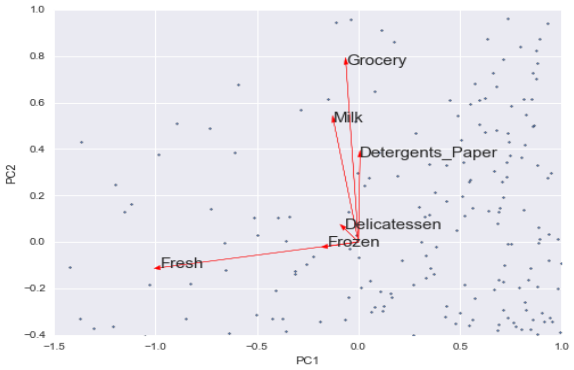

Figure 2: PCA individual component vectors visualized

Shifting focus to the component vectors, their values can be used to support this claim. Both first and second Principle Components can be seen in Table 1 below.

	FRESH	MILK	GROCERY	FROZEN	DET_PAPER	DELICATESSEN
PC1	-0.9765368	-0.1211840	-0.0615403	-0.1523646	0.0070541	-0.0681047
PC2	-0.11061386	0.51580216	0.76460638	-0.01872345	0.36535076	0.05707921

Table 1: PC1 and PC2 component vectors for the dataset

Interpretation of these values gives insight into which features are most strongly correlated to each Principle Component. The values that are further away from zero, in either the positive or negative direction, are the driving features of that particular PC. Examining the first Principle Component, it's clear that there is high correlation to the purchase of Fresh items with its associated value of `-0.9765`. All the other features however, have little influence on this PC based on their relatively low 'weight' values. This can be verified by the Bi-Plot visualization (Figure 2) as the Fresh vector is really the only one highly correlated to the PC1 direction.

The same analysis can be applied to the second Principle Component. The 3 highest values and driving features to this Principle Component are Grocery, Milk and Detergents/Paper with values of `0.7646`, `0.5158` and `0.3654`, respectively. Again, referring back to the Bi-Plot, this is supported as those are the vectors with the largest magnitudes in the PC2 direction. Additionally, since PCAs main function is to identify the directions with the most variance, an associated observation is that both Fresh and Grocery items have the highest standard deviation amongst the features in the dataset. Therefore, it's no surprise that these two features have the largest values associated to PC1 and PC2, respectively.

1.4 ICA Components

As mentioned previously, ICA identifies the individual driving sources for the given dataset. In this particular instance ICA can help identify the different type of customers based on the types of items these customers are purchasing. All ICA components were plotted via an annotated heat map to gain a better understanding of these individual driving sources (Figure 3).

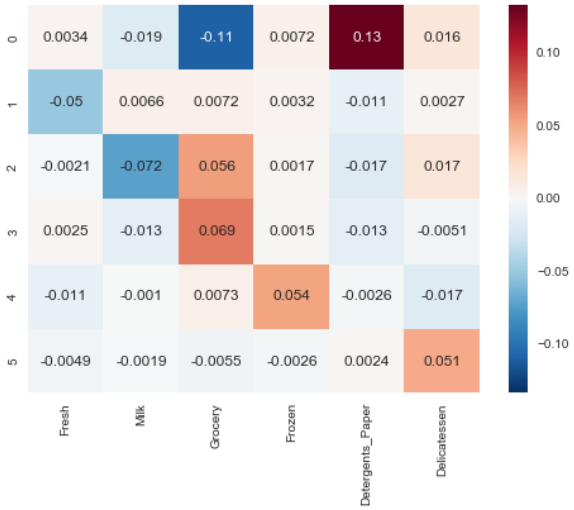

Figure 3: A heat map plot of the ICA components for this dataset.

Conclusions for each of these components are outlined below:

1. The first component vector of this ICA indicates a particular customer that purchases Grocery items and Detergent/Paper items inversely proportionate from each other. Meaning that they purchase almost equal amounts of those items but always at different times.

2. The second vector represents a customer who primarily purchases Fresh items, this could be a store that's main business is selling produce.

3. The third component vector has inverse purchase cycles for Milk and Grocery items which could just indicate that the stocking of these items are on different days of the week.

4. The fourth component of the ICA is primarily buying grocery items and would indicate this customer is likely a smaller general purpose grocery store.

5. The fifth vector indicates this particular customer is typical purchasing frozen items and would likely be a specialty type store.

6. Lastly, the sixth ICA component indicates this particular customer type buys mostly delicatessen items which indicate it would most likely be similar to a New Seasons type of store with a larger deli and prepared foods section.

Since ICA is primarily used for separating superimposed signals, these new independent component sources can be used to identify the particular driving sources (customers) the wholesaler has as linear combinations of the original features.

2 Clustering

2.1 K-Means vs Gaussian Mixture Methods

K-Means is a very simple and effective clustering algorithm that randomly assigns locations to k-centers in the feature space. It then calculates the distances between these centers and each point, and assigns the point to the center that it is closest to. Once all the points have been assigned to an associated cluster center, the algorithm then moves each center to a new location by minimizing all the average distances and essentially locating the new barycenter for those assigned points. This is an iterative process as moving the centers may reassign points to different cluster centers lessening that point's particular "pull" on the location of the center. Once the movement of these cluster centers has stabilized the algorithm ends. However, with K-Means it is good practice to also iterate over the entire process as this will avoid potential problems with local minima since the starting locations of the cluster centers are usually chosen randomly.

> *Advantages of K-Means:* Due to its simplicity, K-Means is easy to understand, it's computationally less expensive than other more complex clustering algorithms and it works well for quickly visualizing potential clusters in a given dataset.
>
> *Disadvantages of K-Means:* K-Means uses hard-assignment for clustering; this indicates that all the points regardless of where they are in the feature space belong to one of the clusters. Therefore cluster centers can be affected by noise or individual and groups of outliers. In addition the algorithm requires a predetermined k-value and as mentioned previously has issues with local minima depending on these initial cluster center locations.

Gaussian Mixture Models (GMM) use statistical probability to generalize K-Means and soft-cluster the data points to associated centroids. The model assumes that each population is generated from a combination of a finite number of sub-populations with unknown parameters.

> *Advantages of GMM:* GMM can be extremely powerful given the fact that not only will it assign data points to individual cluster centers it also computes the probability value that those points belong to those clusters. Also, because it only maximizes likelihood, this algorithm is agnostic to structures and cluster sizes that only 'might' apply.
>
> *Disadvantages of GMM:* Due to its complexity and need to calculate the probability that each point belongs to a cluster GMM can be computationally more expensive than K-Means. It will also always use all the components supplied requiring the need for a secondary technique for deciding the optimal number.

For this particular dataset, GMM will provide a better idea and visualization of the specific customer segments located inside of it. This can be primarily contributed to the fact that its soft-clustering approach will be beneficial in this dataset as the points are fairly spread out around the edges.

2.2 GMM Model Parameter Selection

For the selection of the number of components (clusters) used for the GMM an information-theoretic criterion (BIC) was used. Adopting a technique that was presented on the SK-Learn website, both the n_components (number of components) and covariance_type where supplied as iteratives containing the following values `[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]` and `['full', 'spherical', 'tied', 'diag']`, respectively. Each combination of these values was then supplied to the GMM model and the associated BIC (Bayesian Information Criterion) value was recorded. The scores for the entire set of these combinations can be seen in Figure 4. The model with the combination that returned the lowest BIC score, and is classified as the best model for the data, was the model that had n_components equal to 8 and covariance type of 'diag'.

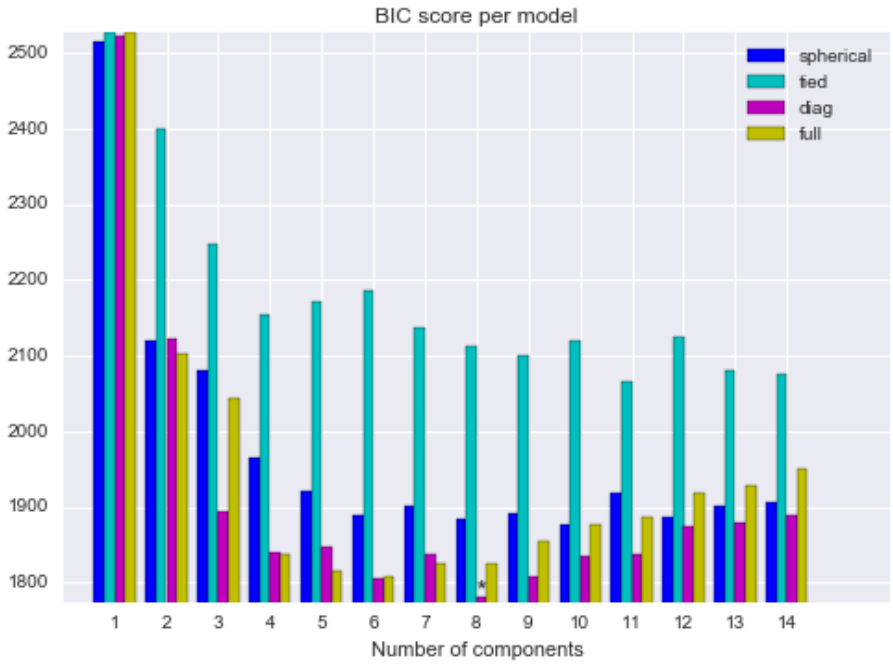

Figure 4: The BIC score per GMM model.

2.3 Fitting the GMM model

In order to visualize the model, PCA was used to reduce the number of dimensions to 2 so that the clusters could easily be plotted and shown in a 2Dimensional space. This would seem somewhat questionable if this dataset didn't reduce as cleanly as it does with PCA. However, as mentioned previously ~86% of the information is retained during this dimensionality reduction. The visualization of the GMM model clusters with 8 components can be seen in Figure 5. It's important to note that the first Principle Component is along the x-direction and the second Principle Component is in the y-direction.

Figure 5: 8 Clusters on the PCA reduced dataset.

This visualization provides relatively well distinguished clusters, however they may be a bit granular. Referring back to the information theoretical criteria graph (Figure 4), rerunning the model fitting and visualization while balancing a smaller number of components and managing the associated increase in BIC score, might provide more distinct and generalized cluster segments.

2.4 Cluster/Customer Types

For each of these clusters the total number of customers associated and the averages of their purchased were parsed from the data set and supplied to Table 2. Also, the key values have been identified and highlighted in red.

Cluster	# of Customers	Fresh	Milk	Grocery	Frozen	Detergents/Paper	Delicatessen
0	84	2245	**7649**	**12195**	1281	5322	1227
1	27	**42668**	2601	3800	5411	718	2029
2	123	**12676**	1596	2011	3660	350	799
3	9	**25043**	**37154**	**47115**	6627	**20716**	9478
4	86	4346	2738	3289	2247	706	1063
5	34	7466	**16039**	**22498**	2284	9986	1965
6	1	**112151**	**29627**	**18148**	**16745**	4948	8550
7	76	**18619**	6530	8342	3951	2192	2138

Table 2: The individual cluster statistics for the dataset and key values can be seen above.

Each of these clusters represents a specific customer type. By focusing on these key values they can reveal insights and one can formulate possible hypothesis on the various customer segments. Each of the clusters and their possible customer types are described below:

Cluster 0: This customer type buys relatively high volumes of Milk and Grocery items, but does not have typically stock Fresh or Frozen items. It's also important to note that while Detergents and Paper was not flagged as a key high value it is purchased in substantial volumes compared to the other categories. This customer could represent a Gas Station or Convenient store they buy some of everything but limit the purchase of Fresh, Frozen and Delicatessen items.

Cluster 1: The major stand out feature for this particular customer type is Fresh items. This is likely a large farmers market or a bigger grocery store that mainly focuses on selling higher volumes of fresh items/produce.

Cluster 2: This cluster is similar to *Cluster 1*, however the volumes are roughly ¼ the scale. This could be a smaller farmers market or a specialty produce store that focuses on quality over quantity.

Cluster 3: Is a one-stop shop type customer (Fred Meyers). They have Groceries, Milk and Fresh items, but you can also find your stationary and house cleaners and stop by the deli on the way out to grab a freshly made sandwich or potato salad.

Cluster 4: These customers are likely family run grocery stores, small volumes but they are still buying items from each of the categories.

Cluster 5: This is an Albertson's type of establishment. They are mainly focused on milk and groceries with less emphasis on being a one-stop shop. However, they do still have those general items if a customer thinks of it while doing some shopping.

Cluster 6: This customer is alone in its own cluster and it's mostly due to the super high volumes of fresh items. This is almost like another wholesaler of produce items or could be a distributor warehouse that supplies a chain of smaller grocery stores.

Cluster 7: These customers are similar to those from *Cluster 2*, in the sense that they value selling fresh items but they still have relatively high volumes of the other items as well. These customers are likely like New Seasons or Whole Foods, where the emphasis is on fresh ingredients, but you can still by general items there as well.

3 Conclusions

3.1 Which technique felt like it fit naturally with the data?

PCA seemed to be the technique used that best fit this particular dataset. The fact that almost 90% of the data's variance could be explained by 2 composite variables was extremely helpful not only for dimensionality reduction but also the customer cluster visualization.

3.2 How would you use that technique to assist if the company conducted an experiment?

According to the original problem statement, this supplier of this wholesale dataset wants to understand their customer base so that if changes are implemented they can have some understanding of how each of these customer segments will be affected. As previously stated, PCA allows for the identification of the directions of maximum variance in the dataset. Or in other words it compresses the data into composite features that represent the essences of the original data. PCA alone can help identify which features contribute most to these composite features and because those contain the most variance they are features that provide information on specific items purchased by all customer types.

Using PCA in tandem with a clustering technique can then not only help identify customer segments, but can also help us visualize them in a realm that has small enough dimensions that we can grasp and make sense of. By understanding the specific customer segments experiments can be targeted to test samples from each of these customer types and truly understand how the changes affect the customer base as a whole.

An example of this is these newly formed clusters can act as decision boundaries on where to split customer data for testing. By utilizing the information that customers inside clusters are similar and by applying A/B testing methodologies, changes can be exposed to only a portion of the customers from cluster in question. If the change is positive, the wholesaler can apply this change to the entire cluster. However, if the change is negative, this change will likely not be well received by similar customers or even similar clusters.

3.3 How would you use that data to predict future customer needs?

Based on the customer segments that have been identified we now understand the types of customers this wholesaler has. It is this level of detail and understanding of who the customer is and how best to cater to their needs that assists in also identifying a solution to the original problem. Customers who build their businesses on marketing themselves as a specialty stores that deliver high quality 'fresh' products may need and want shipments in the morning. This will allow their customers to see that these products are delivered fresh and stocked daily. On the other hand, big box stores or customers who frequently buy frozen products and general groceries may not mind stocking products at night when the store is less busy. These types of things are what are often revealed during A/B testing.

This data could also be used to target specific customer segments by providing them new product advertisements or information in categories that they frequently purchase. This could potentially boost

sales for the wholesaler or just provide a general feeling that the company cares about their success and wants to grow together.

Finally, these clusters can be labeled and fed into a supervised learning model. This model can then be used to predict which cluster new customers will belong to. Equipped with this information their needs can be targeted based on the wholesaler's experience serving past customers that belong to the same associated customer type. This technique could be used for immediate damage control as well. Since the wholesaler is already receiving complaints, those customer's data can also be mapped to a cluster to see if the complaints are coming from similar or a set of similar customers based on the density of the complaining customers in a particular cluster. If that's the case that entire cluster would probably benefit from reinstating the morning deliveries.