Organization

Title

Machine Learning NanoDegree

CREATING CUSTOMER SEGMENTS

AN UNSUPERVISED LEARNING PROJECT

---

by Omoju Miller

UDACITY

August 7, 2016

INTRODUCTION

PROJECT OVERVIEW

This project applies unsupervised learning techniques on product spending data collected for customers of a wholesale distributor in Lisbon, Portugal to identify customer segments hidden in the data.

PROBLEM STATEMENT

A wholesale distributor recently tested a change to their delivery method for some customers, by moving from a morning delivery service five days a week to a cheaper evening delivery service three days a week. Initial testing did not discover any significant unsatisfactory results, so they implemented the cheaper option for all customers. Almost immediately, the distributor began getting complaints about the delivery service change and customers were canceling deliveries–losing the distributor more money than what was being saved.

The goal of the wholesale distributor is to find what types of customers they have to help them make better, more informed business decisions in the future.

The dataset contains 440 data points representing clients of a wholesale distributor. It includes the annual spending in monetary units on the following product categories:

- Fresh
- Milk
- Grocery
- Frozen
- Detergents_Paper
- Delicatessen

To find out what types of customers the wholesale distributors have we suggest the following strategy:

(a) Preprocess the data by applying appropriate techniques like feature scaling and outlier detection.

(b) Explore the data to determine relevant features.

(c) Perform a principal component analysis on the data to understand which features seem to trend together.

(d) Implement clustering to find hidden patterns in a dataset.

For this problem, we don't know how many "segments" of customers there are. In fact, this is what we are trying to learn. As such, we can use the *silhouette score* of each data point to determine how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar). The *mean silhouette coefficient score* can the serve as a metric that can guides us in learning the optimal number of segments in the data.

ANALYSIS

DATA EXPLORATION

The dataset used in this project was created by UC Irvine's' Center for Machine Learning and Intelligent Systems. The data is stored in a CSV file where each row corresponds to an order.

SAMPLE DATA POINTS

To gain a better understanding of the customers and how their data will transform through the analysis, we randomly selected three data points that varied significantly from each other.

Table 0.1: Samples' category spending.

Sample	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	12669	9656	7561	214	2674	1338
1	3067	13240	23127	3941	9959	731
2	32717	16784	13626	60869	1272	5609

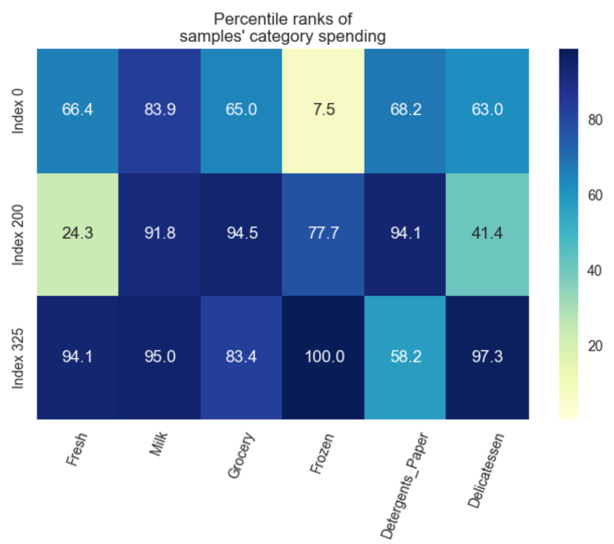

Figure 0.1: **Percentile ranks of samples' category spending.**

As a means of understanding and validating the usefulness of the learned customer segments, we predicted the kind of establishment (customer) each of the three samples represent.

- **Sample 0**

  From the heat-map in figure ?? we can see that this sample represents an establishment that spends very little on textttFrozen category as compared to other product categories. Furthermore, this kind of establishment also spends significantly on textttMilk, at eighty-three percentile to be exact. My intuition leads me to think that this *might* be a **market**.

- **Sample 1**
  This kind of establishment spends less on textttDelicatessen at forty-first percentile and even significantly less on the textttFresh category. On the other categories, their spending is at well over the seventieth percentile range. My intuition leads me to believe this might be a place that serves meals and offers lodgings like a **hotel** or a **university** with a residence halls and adjoining cafeteria.

- **Sample 2**
  This establishment spends significantly on all categories with the exception of the textttDetergents_Paper category where their spending is limited to the fifty-eight percentile. My guess is that this sample represents **wholesale retailer** like a “costco" or “Sam’s club".

FEATURE RELEVANCE

Let’s figure out if it’s possible to determine whether customers purchasing some amount of one category of products will necessarily purchase some proportional amount of another category of products?

To do this, we can train a supervised regression learner on a subset of the data with one feature removed, and then score how well that model can predict the removed feature.

We predicted the impact of the feature `Grocery` on customers’ spending habits. The prediction score was 0.682 out of 1. This score which tends towards being high indicates that it this feature is not necessary for identifying customers’ spending habits. The reason is as follows:

- when customers spend on `Grocery` they also spend on other features, which means we can’t decipher spending habits.

What would be better is to regress on a feature that has a **low** R^2 score indicating low correlation.

To get a better understanding of the dataset, we can construct a scatter matrix of each of the six product features present in the data. From figure ?? we can see that the data for these features is not normally distributed. Its skewed to the right and has a long tail.

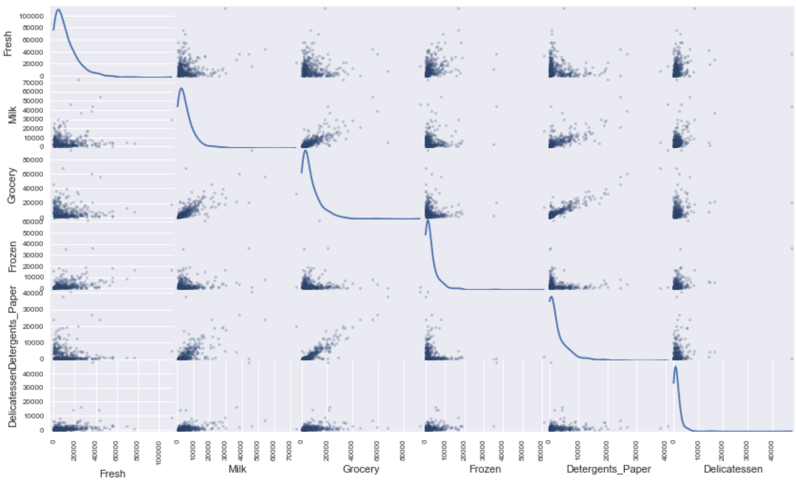

Figure 0.2: **A scatter matrix for each pair of features in the data.**

The pairs `Grocery` and `Detergents_Paper` have a very strong correlation. From the scatterplot, one can see that the data for these two features looks like it could fit a nice diagonal. To formalize my intuition, I performed a pearson's correlation on these two features and got a 0.92 score, noting that a score of 1 is the highest a correlation can be.

Another pair was `Grocery` and `Milk`. From the scatterplot, you can kind of see a trending along the diagonal. It had a correlation score of 0.73.

Another pair was `Milk` and `Detergents_Paper`, the pair had a correlation score of 0.66. This confirms my suspicions that `Grocery` is **not** a strong predictor of overall client spending habits.

OUTLIER DETECTION

The algorithm that we will be using to cluster the data points, `KMeans` is *very sensitive* to outliers. Since this algorithm uses the distance from the centroids of the clusters, i.e. the means to calculate which data points belong to what cluster, leaving values that are very far off from the rest will influence the structure, leading to undesirable clusters.

As part of the data exploration process, we were careful to analyze the data for potential outliers. We used Tukey's Method for identfying outliers to clean the data. We removed 42 data points from our dataset which corresponds to 10% of the dataset. In this case we still have around 390 uniques data points for a problem with 6 variables.

FEATURE TRANSFORMATION

We scaled the data to a more normal distribution and removed outliers. We then applied PCA to the data to discover which dimensions about the data best maximize the variance of features involved.

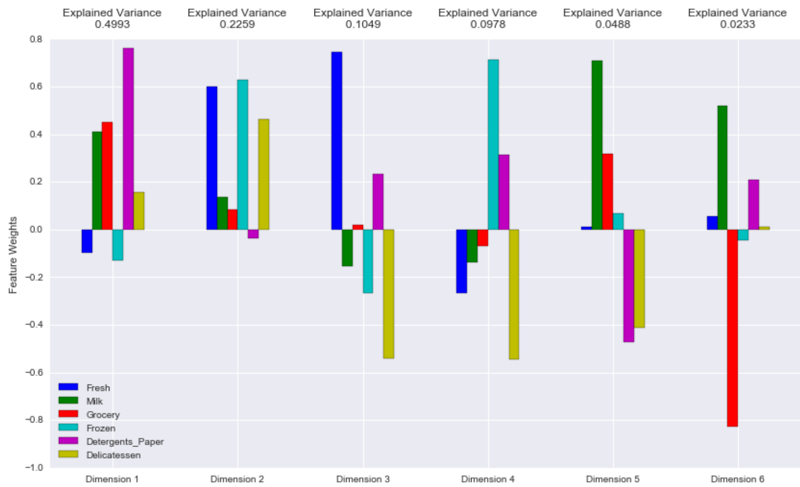

Figure 0.3: **PCA on the log transformed data.**

Unsurprisingly, dimensions 1 and 2 account for almost 72% of the variance in the data. One of the interesting aspects of PCA is that the most interesting dynamics occur only in the first k dimension, and then they fall off. In this case, I would say the first 3 components highlight some clear customer segments.

The first four principal components account for 93% of the variance in the dataset. For this analysis, we have decided that a correlation value above 0.5+- is deemed significant.

- First Principal Component Analysis - Dimension 1
  The first principal component is made up of large positive weights in `Detergents_Paper`, and lesser but still sizable positive weights on `Grocery` and `Milk`. It also correlates with a decrease in `Fresh` and `Frozen`. This might represent spending in household staples products that are purchased together.

  It is important to note that we have a lot of variance in customer spending in the correlated features of `Detergents_Paper`, `Grocery` and `Milk` – some customers buy more while other customers buy less in these three categories.

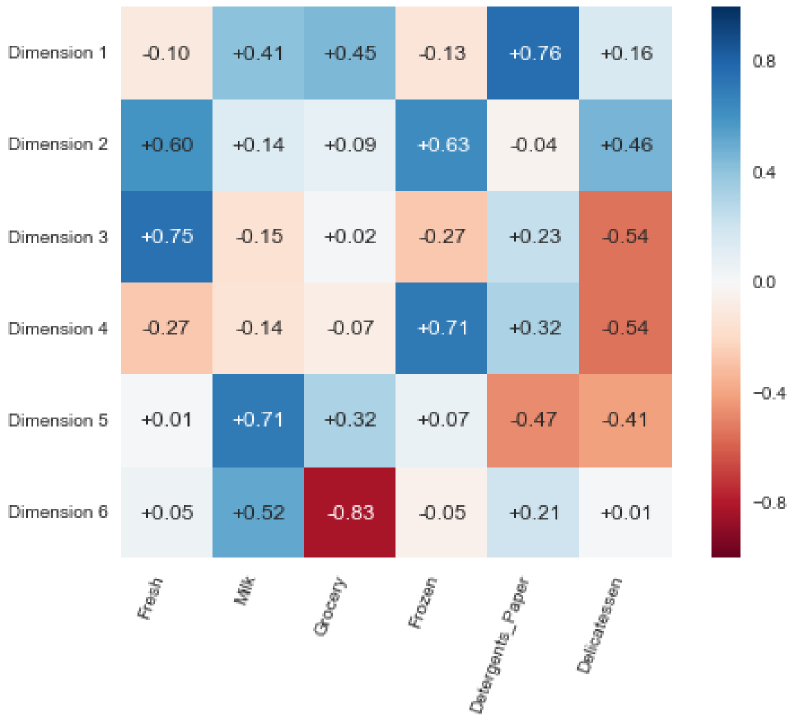

Figure 0.4: **Correlations on the PCA log transformed data.**

- Second Principal Component Analysis - Dimension 2
  The second principal component is made up of large positive weights in Fresh and Frozen, a lesser but still sizable weight on Delicatessen, and smaller weights on Milk and Grocery. It also correlates with a slight decrease in Detergent_Paper. This might represent spending in food items that are purchased together.

- Third Principal Component Analysis - Dimension 3
  The third principal component is made up of a large positive weight in Fresh, a small positive weight in Detergents_Paper, and a significantly lesser positive weight in Grocery. Also, the third principal component has a large negative weight in Delicatessen and a lesser negative weights in Milk and Frozen categories. This might represent spending more in the Fresh category while spending less in Delicatessen category.

  We have some variance in customer spending in the correlated features of Fresh, Detergents_Paper and Grocery – some customers buy more while other customers buy less in these 3 categories.

- Fourth Principal Component Analysis - Dimension 4
  The fourth principal component is made up of a large positive weight in Frozen and a lesser but somewhat sizable positive weight in Detergents_Paper. Also, this principal compo-

nent has large negative weight in `Delicatessen`, a smaller negative weight in `Fresh`, `Milk`, and a significantly smaller negative weight in `Grocery`. This might represent spending more on non-perishable products while simultaneously spending less on perishable products.

The first dimension seems to represent customers who spend *significantly* on `Detergents_Paper`, spend okay on `Milk` and `Grocery`, and none to less on `Fresh` produce and `Frozen` products. Another way to understand these first four dimensions is that they each represent a *customer segment*

IMPLEMENTATION

Because the first two dimension represent a vast amount of explained variance in the data, we reduced the dataset down to those dimensions. We then performed clustering on the reduced dataset.

For clustering, we selected the **KMeans** algorithm for the following reasons:

- We want to have a clear delineation of who the various customer segments are. KMeans gives hard partitions and will allow for this.
- KMeans scales and is a simple and efficient algorithm unlike the Gaussian Mixture Model which does not scale. If we ever grow the dataset, this difference will save a lot of time in the future.
- Easy to understand and compute
- It's efficient. The time requirement is $O(I * K * m * n)$ as long as the number of clusters is significantly less than m.
  - where I is the number of iterations required for convergence
  - where K is the number of clusters
  - where m the number of points
  - where n is the number of attributes

Another algorithm we considered was the Gaussian Mixture Model. Which allows for soft-classification, i.e., data point can belong to more than one cluster with probabilities.

We calculated the mean silhouette coefficient for the data points and determined that the optimal number of clusters to use in describing the data was two.

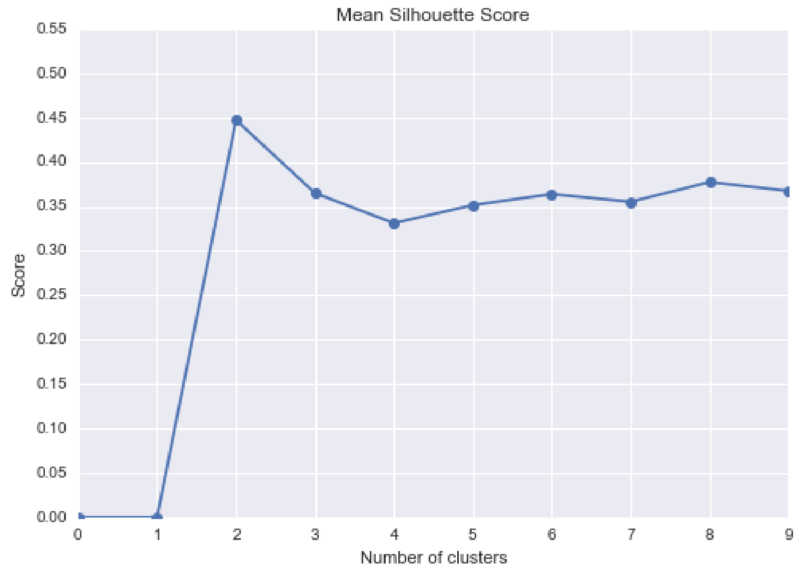

Figure 0.5: **The mean silhouette coefficient score for several hypothetical clusters.**

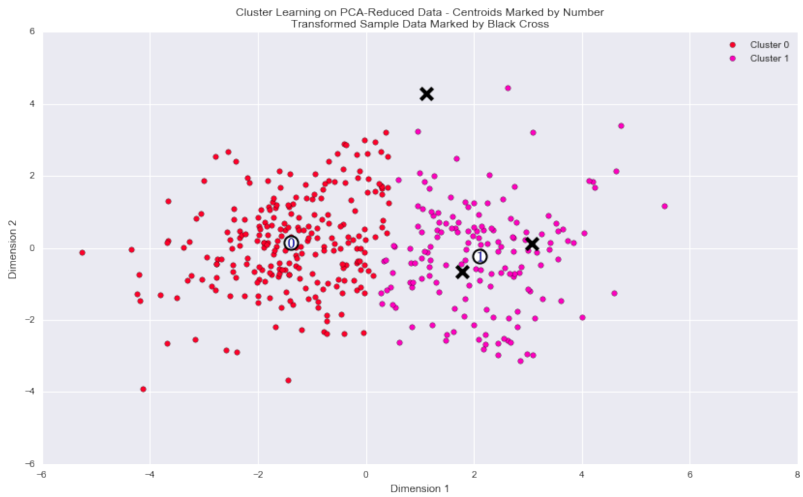

Figure 0.6: **Cluster learning on PCA reduced data.** *Centroids are marked by the numbers, while the transformed sample data are marked by the Xs*

RESULTS

From the two clusters learned from the implementation of the KMeans algorithm, with respect to the problem of creating customer segments, a cluster's center point corresponds to the average customer of that segment. Our goal then becomes describing the kinds of *establishments* that these cluster centers might be describing.

Focusing on **Segment 0**, that cluster center seems to describe establishment that spend significantly on fresh and frozen products. From this, we are inclined to believe that establishment that this segment implies might be best described as belonging to the set that includes, but not limited to **(supermarkets, caterers, restaurants)**.

Also, **Segment 1**'s spending on milk, grocery, frozen and detergents categories is around or over the 75% range for this dataset. Based on intuition, we are led to think that this customer segments might best be described as belonging to the set that includes, but not limited to **(child care centers, public school systems, hotels, hospitals, universities, military bases)**.

Table 0.2: Segments *learned* from clustering.

Sample	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Segment 0	9451	1938	2449	2200	307	771
Segment 1	5424	7780	11532	1123	4444	1136

MODEL EVALUATION AND VALIDATION

For each sample point, we predicted which customer segment best represents it. From our prediction, all three sample points were put in segment 1 as can be seen from figure 7.

Based on our intuition, **Sample 0** doesn't line up. We thought that sample was a small market, and *segment 1* more like a place to caters to feeding and caring for children.

For **Sample 1** our reasoning holds. If we generalize *child-care center* to institutions that feed people and do a lot of laundry like beddings and table cloths, then it makes sense that **Sample 1** is assigned to cluster 1.

The predicted cluster for **Sample 2** doesn't line up. We thought of cluster 1 as a child-care center or restaurant, and **Sample 2** as a wholesale supermarket like Costo. From the clusters in figure ??, we

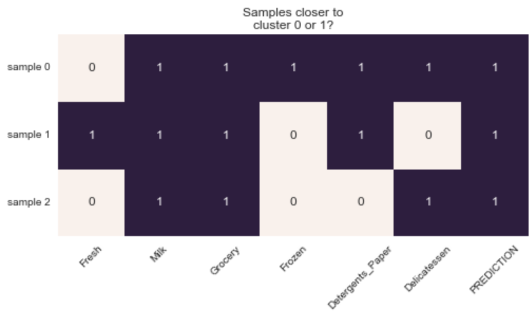

Figure 0.7: **Predicted segments for sample data points.**

can see that this point is amongst the farthest away from the center, which corresponds to a numerically larger data point, so perhaps it still falls in line with our intuition.

justification