Organization

Title

Machine Learning NanoDegree

CREATING CUSTOMER SEGMENTS

AN UNSUPERVISED LEARNING PROJECT

---

by Omoju Miller

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

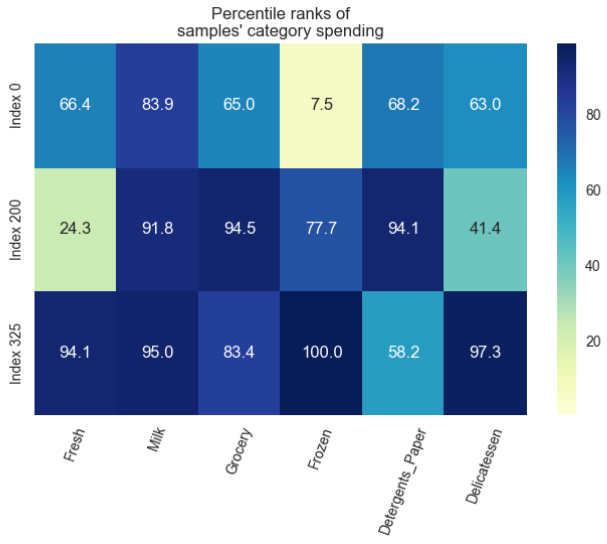
The dataset used in this project was created by UC Irvine's' Center for Machine Learning and Intelligent Systems. The data is stored in a CSV file where each row corresponds to an order.

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



(a)

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

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?

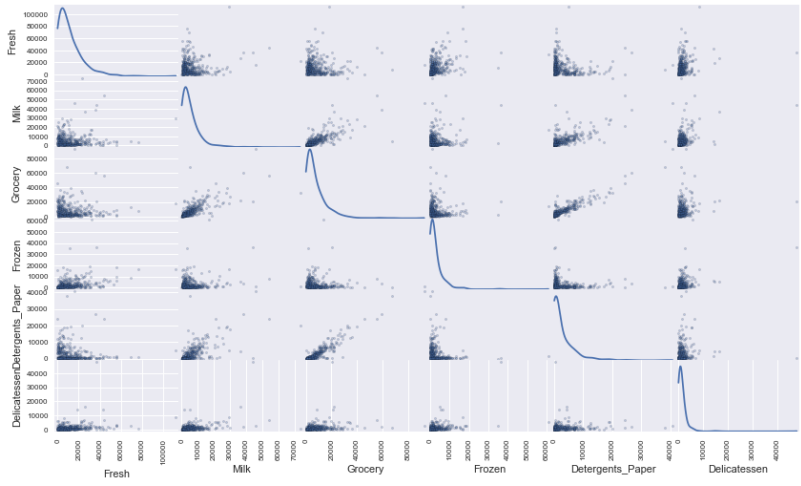
To do this, we can train a supervised regression learner on a subset of the data with one feature removed, and then score how well that model can predict the removed feature.

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 0.2a we can see that the data for these features is not normally distributed. Its skewed to the right and has a long tail.



(a)

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.

EXPLORATORY VISUALIZATION

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

- Have you visualized a relevant characteristic or feature about the dataset or input data?
- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

ALGORITHMS AND TECHNIQUES

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

- Are the algorithms you will use, including any default variables/parameters in the project clearly defined?
- Are the techniques to be used thoroughly discussed and justified?
- Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?

BENCHMARK

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

- Has some result or value been provided that acts as a benchmark for measuring performance?
- Is it clear how this result or value was obtained (whether by data or by hypothesis)?

METHODOLOGY

(approximately 3 - 5 pages)

DATA PREPROCESSING

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

- If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?
- Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?
- If no preprocessing is needed, has it been made clear why?

IMPLEMENTATION

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?
- Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?
- Was there any part of the coding process (e.g., writing complicated functions) that should be documented?

REFINEMENT

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models

to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

- Has an initial solution been found and clearly reported?
- Is the process of improvement clearly documented, such as what techniques were used?
- Are intermediate and final solutions clearly reported as the process is improved?

RESULTS

(approximately 2 - 3 pages)

MODEL EVALUATION AND VALIDATION

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model,Äôs solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

- Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?
- Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?
- Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?
- Can results found from the model be trusted?

JUSTIFICATION

In this section, your model,Äôs final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- Are the final results found stronger than the benchmark result reported earlier?
- Have you thoroughly analyzed and discussed the final solution?
- Is the final solution significant enough to have solved the problem?

CONCLUSION

(approximately 1 - 2 pages)

FREE-FORM VISUALIZATION

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- Have you visualized a relevant or important quality about the problem, dataset, input data, or results?
- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

REFLECTION

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- Have you thoroughly summarized the entire process you used for this project?
- Were there any interesting aspects of the project?
- Were there any difficult aspects of the project?
- Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

IMPROVEMENT

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to

make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- Are there further improvements that could be made on the algorithms or techniques you used in this project?
- Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?
- If you used your final solution as the new benchmark, do you think an even better solution exists?