UDACITY

PROJECT

# Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

## Meets Specifications

Perfect submission!
Exceptional coding work, and analysis demonstrates a pretty fine understanding of clustering in general

Note that I have been a bit lenient at a few places, so please do go through the remarks and the reading material provided to further improve your understanding.

Good luck for the next project!

### Data Exploration

**Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.**

**A prediction score for the removed feature is accurately reported. Justification is made for whether the removed feature is relevant.**

A feature that can be predicted from other features would not really give us much additional information and thus, would be a fit candidate for removal, if we ever need it to make the dataset more manageable.

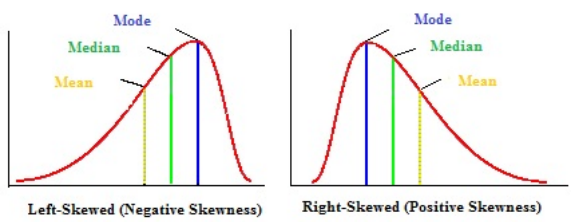
### Miscellaneous remarks:

- Good job fixing the `random_state` while splitting the dataset and for `Regressor` as well, so that we obtain the same score for every run of the program.
- To mitigate the impact of a particular choice of `random_state(s)`, you can average the prediction scores over many values of `random_state(s)`, say, from 0 to 100.

**Student identifies features that are correlated and compares these features to the predicted feature. Student further discusses the data distribution for those features.**

### Remarks:

- The most significant correlation is definitely between `Grocery` and `Detergents_Paper`. `Milk` is also correlated with both these features, but the correlation is relatively mild.
- This complements your interpretation from the previous question. We do get additional information if we keep both of `Grocery` and `Detergents_Paper` in the dataset, but we can drop one just in case we severely need to reduce the dimensionality of our feature space. Later, we will see a better way of reducing the dimensionality of our dataset - PCA.
- Well done remarking that the features' distribution is not normal, but skewed! To be technically precise, the distribution is skewed to the right, as in the following graph:




Clustering algorithms discussed in this project work under the assumption that the data features are (roughly) normally distributed. Significant deviation from zero skewness indicates that we must apply some kind of normalisation to make the features normally distributed.

## Data Preprocessing

**Feature scaling for both the data and the sample data has been properly implemented in code.**

**Student identifies extreme outliers and discusses whether the outliers should be removed. Justification is made for any data points removed.**

### Remarks:

- You have removed all the Tukey outliers, even those for only one feature, effectively removing 10% of samples from our dataset, which is generally not recommended without a strong justification. Therefore, one might choose to remove only the outliers for more than one features, or increase the step size to identify the more extreme outliers.
- Ideally, one must further justify this decision by discussing its impact on PCA and clustering performed later in the project? You might find that your decision here could have a huge impact later on the optimal number of clusters chosen using Silhouette score particularly if you decide to use GMM for clustering.
- You can also check this article for an excellent discussion on this topic, and among the four cases discussed, try to identify which case best characterises the outliers in our dataset.

## Feature Transformation

**The total variance explained for two and four dimensions of the data from PCA is accurately reported. The first four dimensions are interpreted as a representation of customer spending with justification.**

### Remarks:

- Nice work elaborating on each dimension, but it would be better if after remarking the relative weights given to the different features in each dimension, you could surmise what kind of customers might be well-separated along this dimension. For example, a dimension giving relatively high (positive or negative) weights to `Fresh`, `Milk`, `Frozen` and `Delicatessen` would likely separate out the restaurants from the other types of customers.
- Another important point to note is that the sign of a PCA dimension itself is not important, only the relative signs of features forming the PCA dimension are important. In fact, if you run the PCA code again, you might get the PCA dimensions with the signs inversed. For an intuition about this, think about a vector and its negative in 3-D space - both are essentially representing the same direction in space. You might find this exchange informative in this context.

**PCA has been properly implemented and applied to both the scaled data and scaled sample data for the two-dimensional case in code.**

## Clustering

**The Gaussian Mixture Model and K-Means algorithms have been compared in detail. Student's choice of algorithm is justified based on the characteristics of the algorithm and data.**

Good job comparing GMM and KMeans!
From a practical standpoint, the main criteria for deciding between these two algorithms are the speed v/s second order information (confidence levels) desired and the underlying structure of our data.

### Regarding your choice of algorithm:

Your decision to use GMM is perfectly reasonable, particularly since the dataset is quite small and scalability is not an issue.
For large datasets, an alternative strategy could be to go with the faster KMeans for preliminary analysis, and if you later think that the results could be significantly improved, use GMM in the next step while using the cluster assignments and centres obtained from KMeans as the initialisation for GMM. In fact, many implementations of GMM automatically perform this preliminary step for initialisation.

I provide below some citations which might prove useful, if you would like to go deeper into the dynamics of these algorithms:
http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/mixture.html
http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier
http://playwidtech.blogspot.hk/2013/02/k-means-clustering-advantages-and.html
http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm
http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means
http://www.r-bloggers.com/k-means-clustering-is-not-a-free-lunch/
http://www.r-bloggers.com/pca-and-k-means-clustering-of-delta-aircraft/
https://shapeofdata.wordpress.com/2013/07/30/k-means/
http://mlg.eng.cam.ac.uk/tutorials/06/cb.pdf

**Several silhouette scores are accurately reported, and the optimal number of clusters is chosen based on the best reported score. The cluster visualization provided produces the optimal number of clusters based on the clustering algorithm chosen.**

You are absolutely right that Silhouette score should not be the only criterion to decide the optimal number of clusters. But you do need some complementary criterion if you want to choose a number giving sub-optimal score. For example, in this link, 2 is not considered optimal, despite having a better Silhouette score, because it doesn't result in ***balanced*** clusters, while 4 does.

To conclude, you can certainly choose a number giving sub-optimal Silhouette score, but you must justify this choice using some alternate, and **objective**, criterion, not by saying that a higher number would lead to a more *"insightful"* clustering.

**The establishments represented by each customer segment are proposed based on the statistical description of the dataset. The inverse transformation and inverse scaling has been properly implemented and applied to the cluster centers in code.**

**Sample points are correctly identified by customer segment, and the predicted cluster for each sample point is discussed.**

The ideal way of going about this question would be:

- Compare the features of the sample points to those of the `cluster_centers` and thus guess the cluster to which each sample point belong *before* running the code for `predictions`.
- Then, run the code and briefly discuss whether the predictions agree with your intuition or not.
- Lastly, compare to the conjectures made in Question 1 as well.

## Conclusion

**Student correctly identifies how an A/B test can be performed on customers after a change in the wholesale distributor's service.**

Excellent! You have correctly identified the key point here which is to conduct the A/B test on each segment independently, since for an A/B test to be effective, the experiment group (A) has to be highly similar to the control group (B), before the treatment is applied to the experiment group. If they are dissimilar to each other, then the result of the A/B test might be due to some variable other than the variable being tested.

Here are a few links for further reading on A/B testing:
https://www.quora.com/When-should-A-B-testing-not-be-trusted-to-make-decisions/answer/Edwin-Chen-1
http://multithreaded.stitchfix.com/blog/2015/05/26/significant-sample/
http://techblog.netflix.com/2016/04/its-all-about-testing-netflix.html
https://vwo.com/ab-testing/
http://stats.stackexchange.com/questions/192752/clustering-and-a-b-testing

**Student discusses with justification how the clustering data can be used in a supervised learner for new predictions.**

**Comparison is made between customer segments and customer 'Channel' data. Discussion of customer segments being identified by 'Channel' data is provided, including whether this representation is consistent with previous results.**

Remark that the `channel_visualisation` validates, to some extent, the choice of using GMM, as the clusters do have a fair amount of overlap in reality. Although a perfect classification is not possible to achieve, soft clustering gives us confidence levels in our predictions, which would understandably be low at the boundary between two clusters.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

Student FAQ