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.

Excellent work predicting the establishments represented by the sample points based on the comparison of their features to the dataset quartiles!
As we see later, the features' distribution is highly *right-skewed*, therefore, the quartiles definitely serve as a better reference than mean.

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

Your interpretation of the relevance of `Frozen` based on its prediction score is absolutely correct! The low/negative prediction score for a feature means that the values of that feature cannot be predicted well by the other features in the dataset and therefore, the feature is not redundant and may contain useful information not contained in other features.

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. For the exact values, you can use `data.corr()` to get a matrix of correlations for all feature pairs.
- The lack of correlation with `Frozen` nicely aligns with your interpretation of its relevance in the previous question.
- Well done remarking that the features' distribution is not normal, but right-skewed! 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 correctly identified the outliers for more than one features. To make this task easier, you could have used the concept of counter here.
- It is true that the outliers should not be carelessly removed, but there are situations where our analysis could benefit from their exclusion, even when they don't arise from mistake in data entry. Please 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.
- You are still free to keep/remove whatever outliers you like, but you must discuss the impact of including these on the variance of the dataset, and on the PCA and clustering algorithms performed later in the project. In particular, you might find that your decision here could have a huge impact later on the optimal number of clusters chosen using Silhouette score.

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.

Nice work elaborating on the PCA dimensions and interpreting them as a representation of customer spending. Important thing to remark here is that a high/low (absolute) value along the PCA dimension can help differentiate between different types of customers. 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.

The following links might be of some help in the context of this question:
https://onlinecourses.science.psu.edu/stat505/node/54
http://setosa.io/ev/principal-component-analysis/

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.

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.

Good work proposing the establishments represented by the sample points based on the comparison of their features to the dataset mean, but as remarked in Question 1, the median probably serves as a better reference than mean, because of the skewed data distribution.

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

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.

I give below a few links which might help remove misconceptions on this topic, if any:
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.

Good work, and good 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