UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE 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.

Correct, indeed! 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.

Suggestions

- Good job fixing the `random_state` while splitting the dataset. It would nice to do the same 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.
- This is in line with 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.
- Good work remarking that the features' distribution is not normal. You might also want to explicitly mention the skewness of the distribution. 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.

Awesome work with boxcox!

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

Awesome coding work, automating the process of identifying the outliers for more than one features.

You give a good point while justifying your decision to remove these outliers, particularly the impact they might have on PCA and clustering algorithms because of the averaging involved. In our context, PCA and `cluster_centers` turn out to be relatively insensitive to the choice of outliers, but this choice could have a huge impact on the optimal number of clusters chosen using silhouette score, particularly if you use GMM for clustering.

To expand the discussion here, you can also check this article 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.

Good work getting the correct values of cumulative explained variance for two and four dimensions.

Also, nice work elaborating on the first four dimensions and interpreting them as a representation of customer spending. Remark, however, that any PCA dimension, in itself, does not represent a particular type of customer, but a high/low 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
https://www.quora.com/What-is-the-difference-between-K-means-and-the-mixture-model-of-Gaussian

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.

Indeed, `number of clusters = 2` gives the best silhouette score among the many considered!

Important remark regarding the choice of outliers:

This is one place where your choice of outliers plays a huge role. For example, repeat the analysis without removing any outlier. What is the optimal number of clusters that you get?

Miscellaneous remarks:

- From sklearn documentation, the Silhouette Coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. Therefore, it makes sense to use the same distance metric here as the one used in the clustering algorithm. This is `Euclidean` for KMeans (default metric for Silhouette score) and `Mahalanobis` for general GMM.
- For GMM, BIC could sometimes be a better criterion for deciding on the optimal number of clusters, since it takes into account the probability information outputted by GMM. I leave you to experiment with this.

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.

Conclusion

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

You have correctly identified the key point here which is to conduct the A/B test on each segment separately.
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

Student FAQ