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 😄

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!

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 a feature 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.

On the other hand, 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.

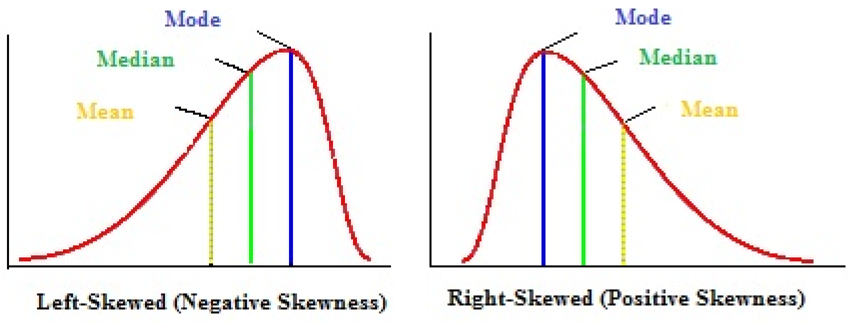
Suggestion:

Your choice of random states can have a huge influence on the R^2-score obtained, which could, in turn, have an influence on your interpretation of the relevance of a feature. To mitigate this, you can average the prediction scores over many iterations, say 100, without setting any of the random states.

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`. This implies that both are weakly relevant as they contain approximately the same information, but if we remove one of them, the other becomes strongly relevant.
- `Milk` is also correlated with both these features, but the correlation is relatively mild. This mildness of correlation explains, to some extent, the low r^2-score obtained for `Milk` in the previous question.
- Correlation for other pairs of features is somewhat insignificant, which also aligns with your interpretation of their relevance in the previous question.
- 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:


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

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:

- Good coding work, correctly identifying the outliers for more than one features. To make this task easier, you could have used the concept of counter here.
- You make an excellent point while justifying your decision to remove all the Tukey outliers, in particular, the impact they might have on clustering algorithms because of the distance averaging involved. In our context, `cluster_centers` turn out to be relatively insensitive to the choice of outliers, unless the outliers end up forming a different cluster by themselves, which could indeed happen if they are not removed at all.
- However, removing all the Tukey outliers, even those for only one feature, effectively removes 10% of samples from our dataset, which is generally not recommended without a strong justification. Taking this into consideration, one might choose to remove only the outliers for more than one features, or increase the step size to identify the more extreme outliers.

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:

- Good coding work computing the cumulative explained variance for the first two and four dimensions!
- 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.
- Also note 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.

The following links might be of interest 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.

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 provided 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.

Excellent analysis!

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, since in A/B testing, everything besides the testing parameter should remain as similar as possible for both the experiment (A) and the control (B) groups, so that we can study the change in behavior caused by the testing parameter.

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.

Excellent discussion, 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.