UDACITY

Logout

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

Meets Specifications

SHARE YOUR ACCOMPLISHMENT

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.

Good work predicting the establishments represented by the sample points based on the comparison of their features to the dataset mean.

As we see later, the features' distribution is highly *right-skewed*, therefore, the median would probably serve as a better reference than mean. In fact, I would recommend comparing to the quartiles to get a better idea of the nature of the establishments represented.

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.

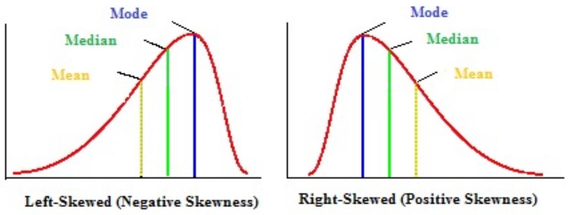
Suggestion:

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.

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.

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.

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, but only by a small margin!

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.
- Silhouette score is not the only criterion to decide the optimal number of clusters. 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.
- 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.

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 analyze each segment independently, although to be technically correct, you must design the A/B tests separately for each segment to begin with. 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.

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