UDACITY

< Return to "Machine Learning Engineer Nanodegree" in the classroom

Creating Customer Segments

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Congratulations on passing this project! :D

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 job here!
Your intuitions are backed up with statistical descriptions of the data!

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

The key takeaway is that a feature which cannot be predicted by other features should not be removed from the dataset for the sake of reducing the dimensionality of our dataset, since its information content, however useful that might eventually prove to be, is not contained in the rest of the features.
You predicted very well the R^2 and the conclusion is great. The attributes with lower R^2 are more relevant since they cannot be predicted by other parameters. The higher R^2 is a less relevant attribute since it doesn't bring any new information to the analysis.

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

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.

Great job discovering the indices of the five data points which are outliers for more than one feature of [65, 66, 75, 128, 154].
Outlier removal is a tender subject, as we definitely don't want to remove too many with this small dataset. But we definitely need to remove some, since outliers can greatly affect distributions, influence a distance based algorithm like clustering and/or PCA! The loss function of the K-means algorithm is defined the terms of sum-of-squared distances, making it sensitive to outliers. In an attempt to reduce the loss function, the algorithm would move a centroid away from the true center of a cluster towards the outlier. This is clearly not the behavior we want.
One cool thing about unsupervised learning is that we could actually run our future analysis with these data points removed and with these data points included and see how the results change.
(http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/)
(http://graphpad.com/guides/prism/6/statistics/index.htm?stat_checklist_identifying_outliers.htm)

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.
Nevertheless, a few comments to improve your understanding:
Strictly speaking, 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.
A corollary of the above remark 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, on running the PCA code again, one might get the PCA dimensions with the signs inversed. For an intuition about this, it is helpful to 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.
Both the algorithms will do fine here, although considering the fact that there are no visually separable clusters in the biplot, one might, indeed, prefer the soft-clustering approach of GMM, 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.

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.

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.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review