UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

4 SPECIFICATIONS REQUIRE CHANGES

Dear Student,

Good work overall. Some changes required, but nothing too painful I hope.

-Gilad

Data Exploration

Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.

Nice discussion of the sample data points. You've compared them to the statistical descriptions of the dataset and used this to inform your suggested establishments.

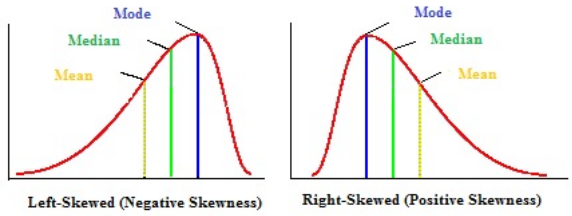
A prediction score for the removed feature is accurately reported. Justification is made for whether the removed feature is relevant.

Also it's worth pointing out that just because a feature may have correlations to other features, doesn't mean it's a good idea to omit that feature. It's worth noting it and investigating further if removing the feature helps or hurts the model. I always base my final decision on feature engineering on the actual performance of the model.

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

Agree that those features are correlated.

The distribution is definitely skewed. You should be able to discuss *how* the data is skewed. Is it left skewed? right skewed?
here is a picture to that helps me remember the terms.



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.

So you need to actually list which points are outliers. Seems silly, but required.

Removing them or not makes little difference for this assignment, but the practice of looking for them is important.
In general when removing outliers I try to ask myself the question. Are these outliers representing samples that might happen again -- i.e. are they rare samples?
Or are they are outliers that probably existed due to human error (reporting error). They will never exist again.

The latter should always be removed. The former -- rare outliers, this requires domain knowledge. The answer to the question "do I want my model to account for them?" will help you decide if you should remove them.

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.

This is a good start. I'm not sure how you are getting that 50% of customers tend to operate this manner or 22% in that manner.
I'm trying to figure out where you got those numbers and it's not clear to me. Perhaps you are using the explained variance number?
the explained variance does not relate to the proportion of customers at all. It relates to how much information from the original 6 features are captured in that dimension.
In general understanding PCA is very hard and hurts most peoples heads. I'll just give some more ways to think about it here.

One intricacy with PCA, that makes it quite difficult to do this kind of analysis is the following:

We *cannot* interpret the sign (+ / -) as representing increased or decreased spending in a particular feature. The signs are actually reversible, and if you run it multiple times on your computer you may have noticed this
http://stats.stackexchange.com/questions/30348/is-it-acceptable-to-reverse-a-sign-of-a-principal-component-score

Remember PCA is a dimension, it's a new "feature", but each store is customer data point is affected by the PCA dimensions differently. How they are affected is what we are looking at.

When doing your analysis I suggest you focus on the absolute values. For example:

When we transform our data, the values for "Dimension 3" show us that...

1. Some customers have a large **positive** 3rd component value — they are likely buying more of the **positive**-weight features while buying less of the **negative**-weight features.
2. There are also customers that have a large **negative** 3rd component value — they buy *less* of the **positive**-weight features while buying *more* of the **negative**-weight features.

PCA has been properly implemented and applied to both the scaled data and scaled sample data for the two-dimensional case in code.

Clustering

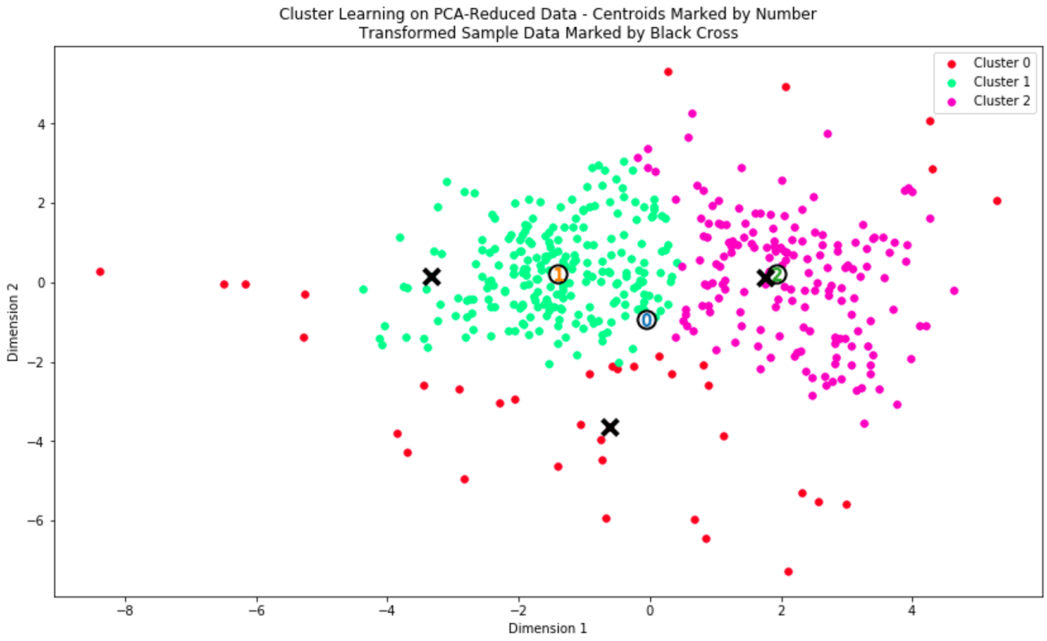
The Gaussian Mixture Model and K-Means algorithms have been compared in detail. Student's choice of algorithm is justified based on the characteristics of the algorithm and data.

Agreed.
In our case our data is one giant blob when projected into a 2d PCA space. This indeed does make GMM a better choice (in my opinion) because
(a) the data is small (so GMM can do it)
(b) it's not clearly defined (so I'd rather have soft clustering).
(c) kmeans will assume a spherical clustering, which we have no evidence of here.

just some extra arguments for you.
Finally keep in mind, often times if the data-set is small like this. I'd just throw everything I have at it and pick what I like best afterwards. It's computationally cheap -- so why not just explore it all?

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.

It's worth pointing out that even though the silhouette score is worse for 3 clusters, if you use GMM you will get a really cool clustering that has an "outlier" cluster. Note that that cluster 0, (red) has points below the center mass and points above (upper right corner has **4 red points**) -- this is basically capturing outliers.



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.

Great discussion, good use of statistics to back it up.

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

Also good discussion

Conclusion

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

This is a good start and understanding. In order to meet specifications you need to explain how we would use the clusters to implement an A/B test. This also includes discussing how to perform the A/B test.
Some resources to help you understand A/B testing

https://www.optimizely.com/ab-testing/
https://en.wikipedia.org/wiki/A/B_testing
https://vwo.com/ab-testing/

Make sure to explain how you would use the clusters to implement an A/B test. how many A groups would we make? how many B groups?

Student discusses with justification how the clustering data can be used in a supervised learner for new predictions.

Yes this is correct. While we could make the output of of clustering algorithm a label for a supervised learner and train the supervised learner to predict it, I always point out -- This is a very silly question!
It would make much more sense to just use the `cluster.predict()` method and always get the right label.

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. Note that while our analysis mostly works out, it's not perfect.
There are a number of overlaps in data points between the two clustering group, such that some data points that will be predicted as Hotel/Restaurant/Cafe customers by our clustering algorithm are in reality Retailers. This misclassification error is less of a concern to this problem since we can now feel confidant our analysis avoids overfitting. In short, we've provided a robust model that has good generalization for unseen data.

This is the reality of all ML algorithms, they get us close -- not perfect. And if it's perfect -- you should be worried!

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Rate this review

Student FAQ