UDACITY

PROJECT

# Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

## Meets Specifications

Hi,

You have done an excellent job of updating the project from last time to meet all the specs here. Even though a few improvements can be made, your overall efforts are commendable. If any of the suggestions are not clear, do feel free to reach out to us on the forums. Happy learning!

### Data Exploration

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

For each sample point, you justify the establishment based on a high or low value of a certain purchase. However, high and low quantities are subjective and unless compared to a certain value, it does not hold much meaning. So, you need to compare the values of each sample point with some central value in the dataset (this can be a dataset statistic like mean, median, etc.) to make an unbiased comparison (like you have done for the first sample point).

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

Great work getting the R^2 score here and commenting on the feature's relevance.

You have correctly observed that the score can change based on the value of the random state used. To mitigate this effect, you should run the decision tree multiple times (you can loop over different values of random states, say 1-100) and average out the score from all the runs to get a more accurate score.

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

Nice job identifying the correlated features and commenting on the distribution of the data. The distribution can also be described as log-normal.

### 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.**

You've done well in identifying the double counted outliers here.

Suggestions

- You could also use a Counter to find these points programatically.
- You have correctly stated that outliers can negatively impact KMeans. Clustering algorithms are sensitive to outliers and algorithms like KMeans give a lot of weight to outliers (as it tries to optimize the sum of squares).
  This paper on the impact of outlier removal on KMeans would be a good read on the topic:
  http://www.math.uconn.edu/~gan/ggpaper/gan2017kmor.pdf

## 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 calculating the cumulative variance and interpreting each dimensions individually. Your interpretation of the negative weighted features is not correct though.

Note here that feature weights are interchangeable (ref.) (the signs of the PCA weights in a dimension can be flipped on running the code again). So a large weight in any direction (positive or negative) will indicate a higher purchase value for the feature. If two features are opposite in sign (in a dimension), it means that they are inversely correlated, meaning that spending on one increases as that on the other decreases. Both the of these features are important in determining customer spending for the dimension. Similarly, a low feature weight (in positive or negative direction) indicates that the customer buys lesser from this feature.

You can read on some examples of interpreting PCA dimensions from the following links:
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.**

Some key differences between the models:

**Speed/Scalability**

- K-Means faster and more scalable
- GMM slower due to using information about the data distribution — e.g., probabilities of points belonging to clusters.

**Cluster assignment**

- K-Means hard assignment of points to cluster (assumes symmetrical spherical shapes)
- GMM soft assignment gives more information such as probabilities (assumes elliptical shape)

You can read more on the differences between the two models here:
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.**

The optimal cluster size has been determined by comparing the silhouette scores for different cluster sizes.

**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.**

Nice job determining the establishments by looking at the dataset statistics.

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

Good work comparing the sample points with the cluster centers to evaluate the predictions made by the clusterer.

## Conclusion

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

Good job identifying how the A/B test can be run here. The main idea is that we need to run two separate A/B tests for each of the clusters. We can judge the impact of the changed delivery schedule on each cluster, and then combine our findings to implement the change only for individual clusters or both the clusters or none of them.

You can read more on A/B testing from the following links:
https://en.wikipedia.org/wiki/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/

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

You have correctly noted that the created customer segments can be used to turn this into a classification problem.

**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.**

Even though there is some overlap in the central region, the overall alignment is pretty good!

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

Student FAQ