UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

4 SPECIFICATIONS REQUIRE CHANGES

Hello Student,
You almost get it done, keep it UP!
There are only few minor mistakes you need to amend, I believe you already understand most of the concept in this project.

P.S. Please include the necessary file in the zip file only, which are the ipython notebook, html file and image file if necessary.

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 work directly stating what establishment the three different samples might represent. Below are some comments and feedback:

Required:

- As required by the specification, please support your claim of establishment representations based on the **statistical description** of the dataset.
- This means that you can compare the samples of the data chosen to some central values of the dataset.
- Here is an example:

  > For the first sample, I think the customer could be a grocery store as their consumption on Milk, Grocery and Detergents_Paper are much higher than the mean and even reach the top 25%. However, its Fresh, Frozen and Delicateseen consumption are less than the bottom 25%. Which means it is very unlikly that the customer is selling food beverage.

↻

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

Nice job finding the `R_2` of all the features, and commenting on their relevance. Please see my comments below:

Required:

- As required by the question, please choose one particular feature, and comment on whether it is necessary or not in identifying customers' spending habits, as I couldn't see this from your answer in the report.

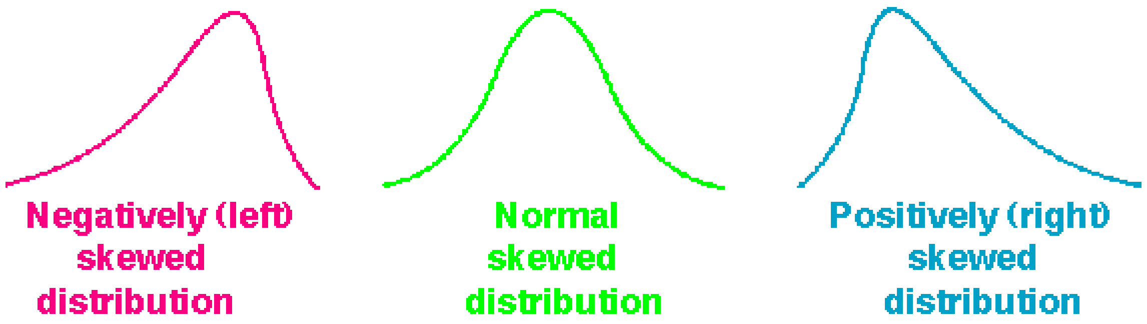
✓

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

You did great identifying the features that have correlation from the dataset.

Suggestions and Comments:

- Kindly note that the distribution is skewed to the right (most points lie to the left of the graph). The graph below interprets how to quickly judge the skewness of a distribution:


Negatively (left) skewed distribution
Normal skewed distribution
Positively (right) skewed distribution

- As you can see from the scatter plot, there are a number of outliers for most of the features.

- Also, the median falls below the mean, and there are a large number of data points near 0.

Data Preprocessing

✓

Feature scaling for both the data and the sample data has been properly implemented in code.

Nice work implementing feature scaling :)

↻

Student identifies extreme outliers and discusses whether the outliers should be removed. Justification is made for any data points removed.

Good try here. Please see my feedback for this section, including what is required:

Required:

- As required by the question and specification, please identify all of the data points which occur as outliers **twice or more**, report them, and choose whether to remove them or not, with justification. You only provided the justification.

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.

Fantastic work in identifying the amount of variance explained by the first 2 and the first 4 dimensions! Please see some feedback and what is required for this section:

Required:

- As required by the question and specification, please interpret **each** of the first to fourth dimensions of the dataset as a representation of customer spending with justification.
- Please reference the absolute feature weights present in the components to justify your answer.
- Please suggest the customer spending of first four dimensions might be representing.
- For example, for a particular dimension, you could say:

  > a significant positive weight is placed on Detergents_Paper with meaningful positive weight on Milk and Grocery. This dimension is best categorized by customer spending on retail goods.

✓

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

Nice job here!

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.

Great work elaborating on GMMs and K-means, and giving a solid reason as to why you chose [K-means] algorithm for this particular problem.
You might want to provide some citations and reference for your work to make it more credible.
Below are some of my comments, feedback, and suggested reading:

Gaussian Mixture Models

- You did well on giving the advantages of Gaussian Mixture Model.

SUGGESTED READING:

- If you feel as going deeper with regards to Gaussian Mixture Models, check out the following links:
  - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/mixture.html
  - http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier
  - http://scikit-learn.org/stable/modules/mixture.html#gmm-classifier

K-Means Clustering

- Your description on the advantages of K-means is very explicit

SUGGESTED READING:

- Check out these links for even more thorough explanations of K-Means Clustering:
  - http://playwidtech.blogspot.hk/2013/02/k-means-clustering-advantages-and.html
  - https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
  - 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

Reason for choice of Algorithm:

- Well done here again! Please see the links below for some more information on how to compare the two algorithms:

SUGGESTED READING:

- 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.

Great job identifying the optimal cluster score as 2, and identifying its associated silhouette score

✓

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.

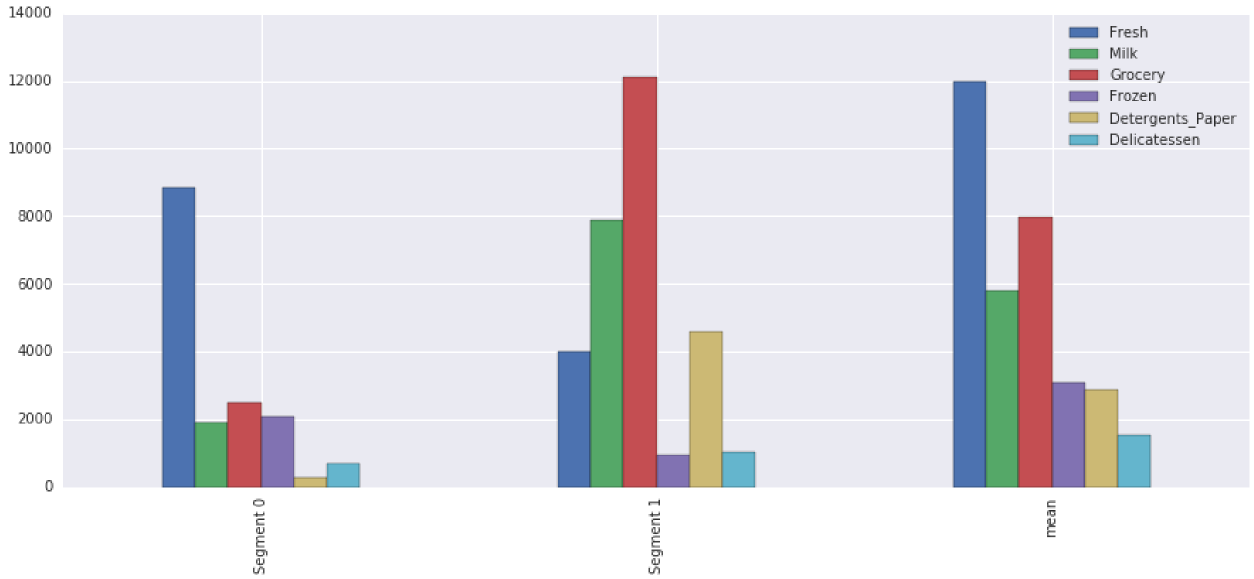
Well done recovering the true centres, and proposing establishments using guidance from the statistical description of the dataset!

Pro Tips:

COMPARING TO DATASET AVERAGE

You could quickly draw a bar plot to visualise the amount of each product purchased for each `true_center`, together with the dataset mean.

```
# Import Seaborn, a very powerful library for Data Visualisation
import seaborn as sns
true_centers = true_centers.append(data.describe().loc['mean'])
_ = true_centers.plot(kind='bar', figsize=(15,6))
```



This will make comparing the three different sample points with each other much easier.

✓

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

Nice work identifying the sample points by customer segment, and discussing the predicted cluster for each sample point.

Conclusion

✓

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

Great explanation here!
The key is to conduct the A/B test on only one segment at the time, rather than on both segments. I like your design of using 2 A/B tests, one per segment.

Suggestions and Comments:

- You can find more information about A/B testing from this link. You can also refer to the Udacity course to get an overview on how A/B testing is conducted (please note that you don't need to take the whole course, just skimming through the lecture videos and overviews should be alright)
- These links were also really helpful to me when picking up A/B testing. Hope they help you too!
  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.

Fantastic reply here
Nice job stating that the cluster labels can be used as an **target variable** to a supervised learning algorithm!

✓

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.

Well done comparing the clusters from your algorithm to the customer 'Channel' data!

RESUBMIT

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review

☆☆☆☆☆