UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

9 SPECIFICATIONS REQUIRE CHANGES

Great start! There are a couple points that need to be altered, but you've got a good foundation to build from. Keep at it - we look forward to the next one.

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 start here, but you'll need to make sure to use statistical descriptions of the dataset in your answer. This means explicitly talking about how the features compare to the median/mean/quantiles of the dataset

TIP

In general, I find it really helpful to visualize sample points when I'm trying to figure out what they represent. You can do this quite simply with the following code 😄

```
import matplotlib.pyplot as plt
import seaborn as sns

samples_for_plot = samples.copy()
samples_for_plot.loc[3] = data.median()
```

```
labels = ['Sample 1','Sample 2','Sample 3','Median']
samples_for_plot.plot(kind='bar')
plt.xticks(range(4),labels)
plt.show()
```

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

Great!
You nailed the key point here - if we can reliably reconstruct a feature from other features, it probably doesn't contain a whole lot of unique information. 👍

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

It doesn't look like you've answered this point:

- How is the data for those features 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.

Your code here isn't quite correct.

```
for feature in log_data.keys():

    # TODO: Calculate Q1 (25th percentile of the data) for the given featu
re
    Q1 = np.percentile(data, 25)

    # TODO: Calculate Q3 (75th percentile of the data) for the given featu
re
    Q3 = np.percentile(data, 75)
```

First, note that you are iterating over each feature one at a time, so make sure you are only getting the percentile of the current feature. Second, make sure you are using the log-transformed data and not the original data.

E.g

```
for feature in log_data.keys():

    # TODO: Calculate Q1 (25th percentile of the data) for the given feature
    Q1 = np.percentile(log_data[<<INSERT VARIABLE>>], 25)

    # TODO: Calculate Q3 (75th percentile of the data) for the given feature
    Q3 = np.percentile(log_data[<<INSERT VARIABLE>>], 75)
```

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.

Awesome start here, we just require that you go a little deeper into the meaning of these components.

> In the first dimension the fresh goods and frozen goods have greater positive effect whereas remaining all show a negative effect on this dimension and this is the first principal component as the explained variance is highest for this.

Based on this description, what sort of customer would score high in this component? What about a high negative score? You should try to figure out which customer segments might be represented by high positive vs. negative scores here. If you have any trouble analyzing the pca components, check out this post which has some explanation as to interpreting them.

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.

Great work!

In general, K-Means offers better performance if we care about

- Speed
- Scaleability
- Simplicity

Whereas GMM provides more

- Flexibility
- Robustness

The fact that K-Means assumes that all clusters is globular is a pretty enormous assumption, and is always something we have to take into consideration. GMM is far less rigid in this - it allows these spheres to be stretched and compressed.

There are a ton of other models you can use that weren't discussed in lectures as well. One of my personal favourites is DBScan, which uses a *density* measure rather than a distance measure to determine clusters. This can allow for far more unrestricted cluster shapes, which makes this algorithm quite powerful!

If you're interested, check out the links below for more on this subject 😄
https://algorithmicthoughts.wordpress.com/2013/05/29/machine-learning-dbscan/
https://www.quora.com/What-is-an-intuitive-explanation-of-DBSCAN
http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

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.

You actually did perfectly fine here - I'm only marking this down as a reminder to double-check it after fixing your outliers issue. It's possible that your optimal number of clusters will change.

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.

> I could not understand what the question is.....

In this section, you need to analyze the purchasing patterns of each of the segments to try to determine what type of customers make up that segment. Let's take the Segment 0 as an example - you can see that it scores quite high relative to the median in Fresh, but lower than the median in most other categories. What sort of establishment do you think this might represent? Try to answer this question for all segments.

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

Please revisit this section after fixing the previous one.

Conclusion

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

Good start here - you've described an intuitive answer to the question. However, what we're looking for here is a description of how to design *an experiment* using A/B testing on our dataset. Before answering this question, make sure you understand:

- What is A/B testing?
- How is A/B testing applied?

Then answer the following points:

- How do we apply A/B testing to this dataset to figure out an optimal delivery schedule?
- Should we apply A/B testing to each segment individually or all at once?

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.

Please revisit this section after fixing the segment analysis (Question 8).

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)

RETURN TO PATH

Rate this review

Student FAQ