UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

11 SPECIFICATIONS REQUIRE CHANGES

Data Exploration

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

Awesome

- Good job selecting samples of the data.

Required

- It seems like what you're doing here is determining the establishments represented by looking at how much is spent on a category over others meaning you're comparing categories to other categories, the actual way to go about this is to determine this by comparing each category to itself, what this means is to see how much is spent on say, Milk compared to the mean or median of milk, this tells us if a particular customer is spending more or less than the general average. The code included below will help you determine this.

```
# Import Seaborn, a very powerful library for Data Visualisation
import seaborn as sns
```

```
samples_bar = samples.append(data.describe().loc['mean'])
samples_bar.index = indices + ['mean']
_ = samples_bar.plot(kind='bar', figsize=(14,6))
```

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

Awesome

- I like that predicted for every feature. Good job!

Required

> Is this feature necessary for identifying customers' spending habits?

- The final question here hasn't really been addressed, you mentioned that it's easier to predict Milk, but what the question is asking is whether or not this feature is relevant for predicting customer's spending habits. To answer consider that a score that can be easily predicted isn't very relevant since it can be predicted by other features, and features that can't are relevant.

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

Awesome

- Great job noticing the correlated features.

Required

> My initial prediction that Milk is correlated with other features such as Detergents paper and Grocery.

- the objection here is actually similar to the previous section, what this question is asking is to reference the discussion on relevance in the previous section, so that would need to be addressed first. The idea here is to reconcile this discussion on correlated features with the relevance or irrelevance (whichever you go with) of the feature dropped in the previous section.

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.

> Answer: Yes these data points are considered outliers according to the above definition. Datapoints should be removed if we consider to implement ICA/PCA as we are aiming to get the general variance of the data, outliers may represent difficult situations for PCA to capture the main variance on an axis (they can be considered as 'noise' to PCA).

Required

- Some of the answers required here haven't been addressed. What are the indices of the outliers for more than one feature? were only these data points removed or were others also?

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

- Good job on the discussion included.

Required

- A few more issues need to be addressed here. The first is that the total variance explained for the first 2 and 4 dimensions should be provided. This can be done by simply adding the first 2, and the first 4. An even better way of doing this would be by -

```
pca_results.cumsum()
```

- Also there should be some interpretation here of customer spending. The answer basically, should look something like this, take for the first dimension -

> 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.

```
pca_samples =pca.transform(log_data)
```

Required

- The pca_samples should contain a transform of the log_samples not the log_data.

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.

Awesome

- Great job on your discussion on GMM, you're absolutely right it's main advantage is it's soft clustering ability.

Required

- I don't understand a lot of what was discussed on K-Means, take for example -

> K-Means works well with numerical attributes.

- All attributes ideally will be either numerical, or categorical turned numerical variables. Is there something I'm missing?

Suggestion

- I would suggest including references or links to the pages, it's always a good idea to cite other works, also it allows the reader to read more on what you discussed and understand it better.

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.

Awesome

- Great job! Multiple cluster numbers attempted for both models with 2 chosen as optimal.

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.

> Answer: The sepperation into these two segments appears to be of two different types of businesses; one majorly dealing with minor amount of goods (non-bulk vendor), and the other representing bulk vendors which sell large quantities of more than just one product.

Required

- The objection here is similar to the very first objection, some sort of summary statistic (mean/median/percentile) should be used in determining what establishments each segment might represent.

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

> Answer: Most points seem to be consistent with the cluster results, high buyers vs low buyers.

Required

- The code here should be commented and run, and a discussion should be included on whether your earlier discussion matches the clusters the points are predicted to be in.

Conclusion

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

Awesome

- Good discussion here.

Required

- This doesn't really address the question asked here. See the link included for an explanation for A/B tests to see how this is generally performed. What this question is asking is how would you stage an A/B test for this problem?

https://vwo.com/ab-testing/

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

Required

- The question hasn't been addressed. What this is asking is how would you use the results of the clustering to implement a supervised learning model? I would suggest reading the question again for a better understanding of what is required. Basically, what would be predicting in this hypothetical supervised learning process? and what would be our features? how would the clustering/clustering results help here?

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.

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

Student FAQ