UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

5 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.

Good work here but please make sure to also compare each of the samples expenditures to the statistical description of the data (e.g. you could compare each sample against the mean expenditure).

Try referencing the normalized sample expenditures:

```
import seaborn as sns

sns.heatmap((samples-data.mean())/data.std(ddof=0), annot=True, cbar=False, square=True)
```

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

> I attempted to predict "Grocery" and did so with a score of .829, which is very high. Since it can be predicted with a high score, potentially it is not necessary for identifying spending habits - it is highly correlated with other spending categories.

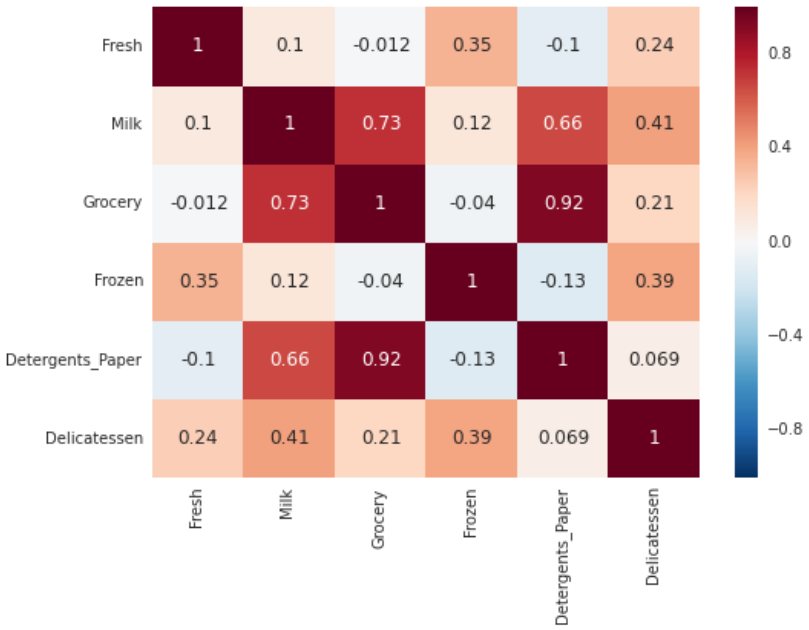
Correct, a high R^2 score here suggests that the feature could be dropped without us loosing too much information.

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

> "Grocery" looks to have a lot of correlation with "Detergents_Paper" as well as with "Milk". I picked "grocery" as it seemed to be a suspiciously broad topic, and my suspicions were confirmed as it shows some degree of obvious correlation with multiple other features.

Correct, we can also visualize pairwise correlations as follows:

```
import seaborn as sns
sns.heatmap(data.corr(), annot=True)
```



> Overall, the data does not look to be normally distributed, and shows a large skew and long tail, with most of the data shifted towards zero. Maybe a log scale would make the data look more normal.

Correct, the feature distributions follow more closely the log-normal distribution.

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.

> 75 is an outlier on multiple features (detergent and grocery) and can be removed.

Correct, please make sure to find and report all other such outliers.

Tip: Try using a counter.

```
from collections import Counter

c = Counter()

for feature in log_data.keys():

    Q1 = np.percentile(log_data[feature], 25)
```

```
    Q3 = np.percentile(log_data[feature], 75)

    step = (Q3 - Q1) * 1.5

    c.update(log_data[~((log_data[feature] >= Q1 - step) & (log_data[feature] <= Q3 + step))].index.value
s)

print [o for o in c.keys() if c[o]>1]
```

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.

> 71.45% of the data is explained by the first and second principal component. 93.06% of the data is explained with the first four principal components.

Correct, the first four principal components capture most of the variance.

> The third PC is weighted heavily on deli and frozen, and negatively weighted fresh.

Correct, a customer scoring high in this component would thus be one buying large quantities of deli and frozen but almost no fresh.

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

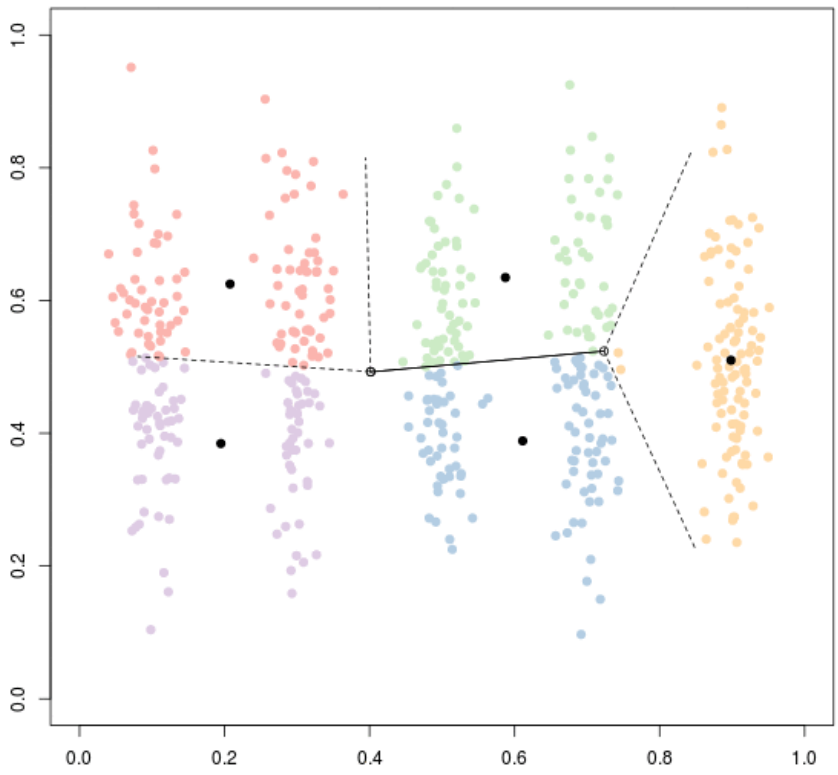
Clustering

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.

> It is fast to run

Correct, perhaps the greatest advantage of K-Means over GMM is that the former is faster and therefore more scalable than the latter.

An additional advantage of GMM over K-means is its capacity for modelling all sorts of elliptical clusters (as opposed to only spherical). This blog post provides a good example of a dataset where GMM would do a much better job than K-means due to the ellongated nature of the clusters.



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 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.

Please make sure to compare each cluster against the statistical description of the data. You could, for example, compare the clusters' mean expenditures against the populations mean expenditures per category.

Try referencing the normalized cluster expenditures:

```
import seaborn as sns

sns.heatmap((true_centers-data.mean())/data.std(ddof=1), annot=True, cbar=False, square=True)
```

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

> sample 0 and 1 are best represented by retailers, and sample 2 by a restaurant/cafe.

Correct, according to the model, the above is the most appropriate categorization but does it make sense? In other words, does each of the samples above actually follow the spending pattern of their assigned segments? You could try having a look at which categories are either above or below the population's mean in each segment and see if the samples above follow the the pattern of their assigned cluster.

Conclusion

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

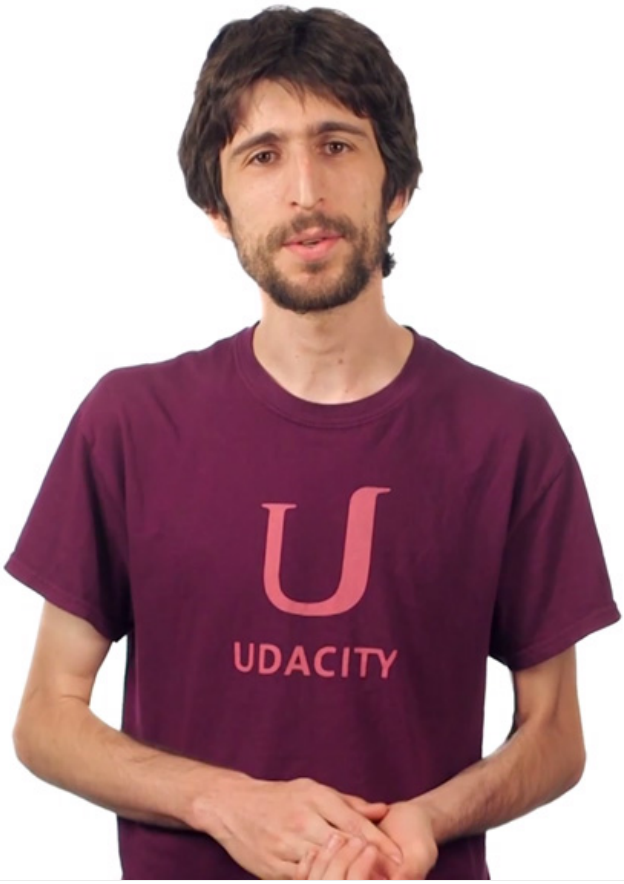
Required: Please make sure to describe how we would actually go ahead and A/B test which segment actually prefers the new service and which does not. In particular, make sure to describe how many A/B tests we need to perform and which customers should be involved in which test.

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.

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)