UDACITY

‹ Return to "Machine Learning Engineer Nanodegree" in the classroom

DISCUSS ON STUDENT HUB

Creating Customer Segments

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

🏆 Terrific job! You seem to have grasped all the main concepts of unsupervised learning and clustering introduced in the project. 😎

Data Exploration

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

💡 **Tip: look at percentile ranks**

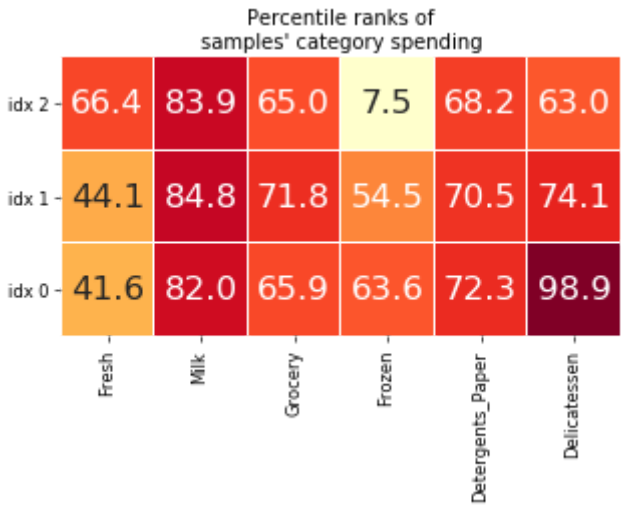
As you can see later in the notebook, the distribution of our customers' spending data has some large skew, so another thing you could try is looking at the category spending percentile ranks with a heatmap:

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

# look at percentile ranks
pcts = 100. * data.rank(axis=0, pct=True).iloc[indices].round(decimals=3)
```

```
# visualize percentiles with heatmap
sns.heatmap(pcts, annot=True, annot_kws={'size':20}, linewidth=.1, vmax=99, f
mt='.1f', cmap='YlOrRd', square=True, cbar=False)

plt.yticks([2.5,1.5,.5], ['idx '+str(x) for x in indices], rotation='horizont
al')
plt.title('Percentile ranks of\nsamples\' category spending');
```



(above shown as an example)

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

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

Data Preprocessing

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

Nice job scaling the data — this will help our data appear more normally distributed and appropriate to use with a variety of machine learning techniques.

You could also simplify the implementation a bit with something like this...

```
log_data = np.log(data)
log_samples = np.log(samples)
```

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

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.

Excellent work reporting the cumulative variance and identifying what the category weights in each dimension represent. 😎

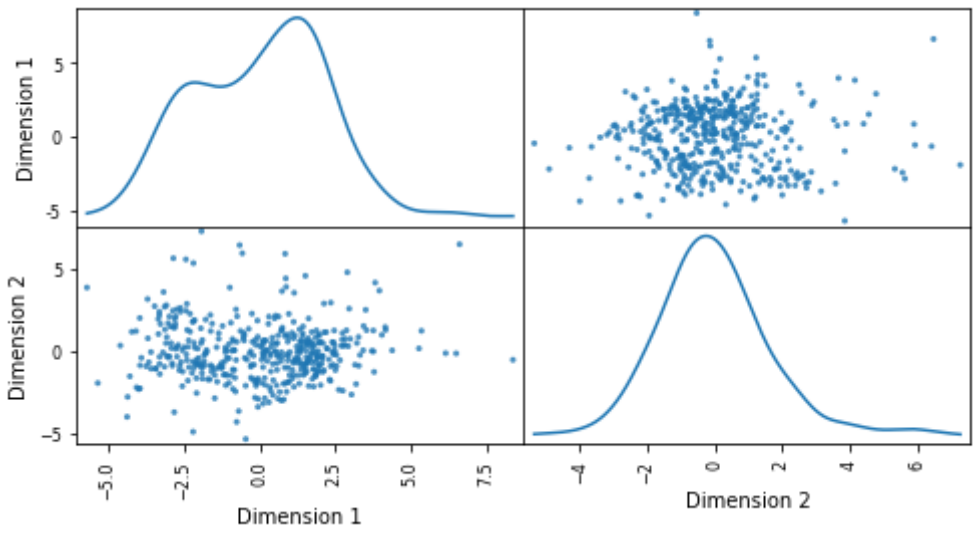
Basically, PCA deals with feature correlations and the variance of the data:

- e.g., the 1st component shows there is a lot of variance in purchases of **Milk, Grocery & Detergents_Paper**
- Customers with **LOWER (negative)** 1st component values (e.g., retailers) purchase a lot of these 3 categories, while those with **HIGHER (positive)** component values (e.g., restaurants) purchase very little.

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

💡 **Tip: look at scatter matrix**
If we view a scatter matrix of the reduced data, we can also see 2 humps in the 1st Dimension that indicate the presence of 2 distinct groups within the distribution...

```
pd.plotting.scatter_matrix(reduced_data, alpha=0.8, figsize=(8,4), diagonal=
'kde');
```

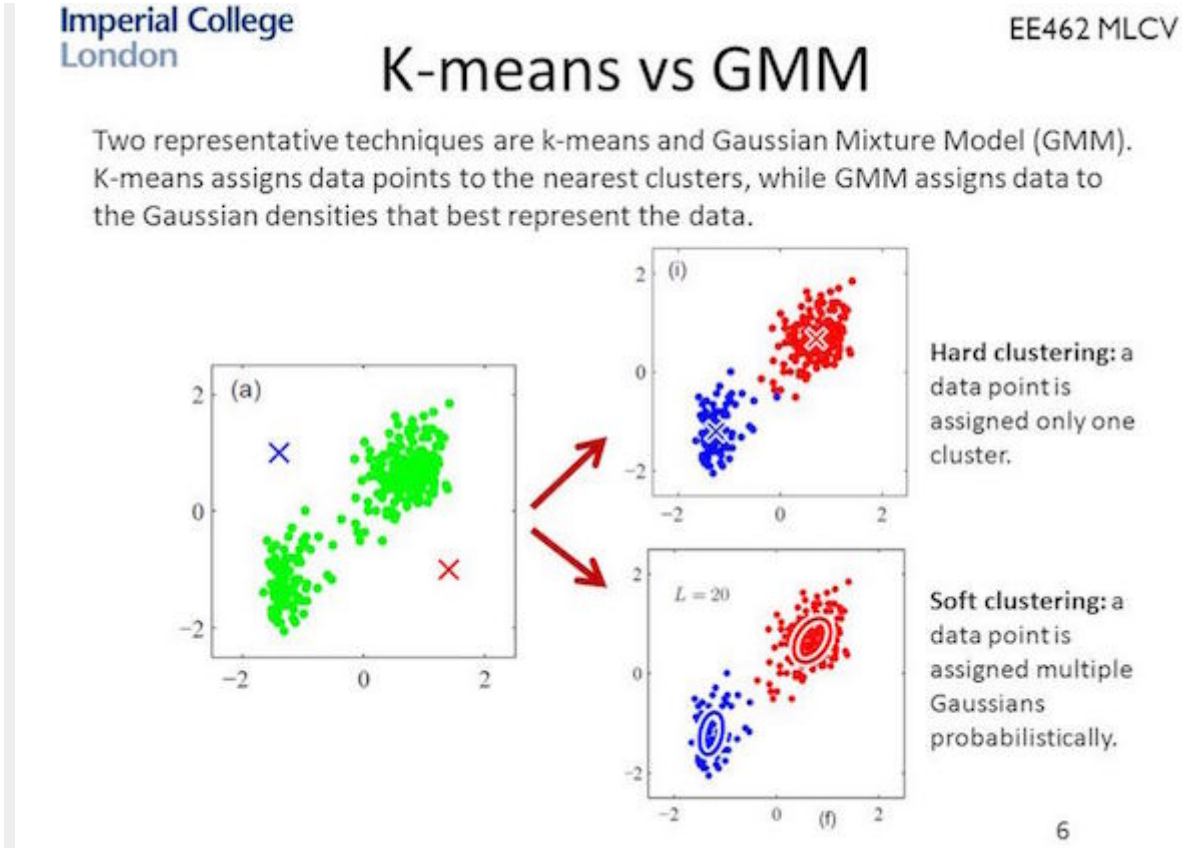


Clustering

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.

💡 **Tip: K-Means vs GMM shape**

Note the big drawback with KMeans is that it assumes the groups are spherical (globular) shapes that are symmetrical, which don't always occur with real data. GMM assumes the groups to be elliptical...



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.

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

Conclusion

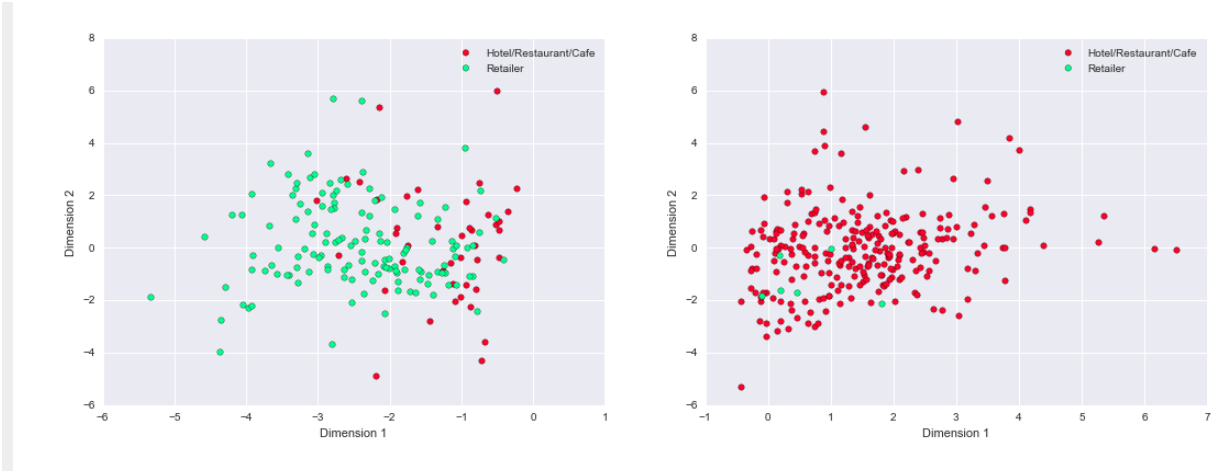
Student correctly identifies how an A/B test can be performed on customers after a change in the wholesale distributor's service.

Good working pointing out how we can test the segments — the customers in each cluster might be affected differently by a delivery change, so to test all the customers properly we essentially need to run multiple separate A/B tests.

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.

Good examination of the 'Channel' data in relation to your learned clustering. As you found out, the 'Channel' data and segments from a **TWO cluster** analysis are pretty well-aligned:



DOWNLOAD PROJECT

RETURN TO PATH

Rate this review