UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

2 SPECIFICATIONS REQUIRE CHANGES

This is a very solid analysis here and impressed with your answers. You have an excellent grasp on these unsupervised learning techniques. You just need to fine tune a couple of these section and you will be good to go, but should be a simple fix and great for learning the material even better. Keep up the great work!!

Data Exploration

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

Excellent justification for your samples here by using the percentile values of the dataset. As using the median/percentiles is much more appropriate than mean, since the median/percentiles are more robust to outliers, which we have here. Great job.

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

> "It helps to predict customers' spending habits as that information can't be predicted from other features."

Spot on! Frozen is an independent feature, so necessary. Thus if we have a high r^2 score(high correlation with other features), this would not be good for identifying customers' spending habits(since the customer would purchase other products along with the one we are predicting, as we could actually derive this feature from the rest of the features). Therefore a negative / low r^2 value would represent the opposite as we could identify the customer's specific behavior just from the one feature.

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

Great job capturing the correlation between features. We could actually get some more insight by looking at numerical correlation by adding it to the plot as well with

```
axes = pd.scatter_matrix(data, alpha = 0.3, figsize = (14,8), diagonal = 'kde')
corr = data.corr().as_matrix()
for i, j in zip(*np.triu_indices_from(axes, k=1)):
    axes[i, j].annotate("%.3f" %corr[i,j], (0.8, 0.8), xycoords='axes fraction', ha='center', va='center')
```

And good ideas regarding the data distributions with your comment of "*The data is not normally distributed, it is skewed (right-skewed).*" Skewed right is great! Could also mention Log Normal. As we can actually get an idea of this from the basic stats of the dataset, since the mean is above the median for all features. We typically see this type of distribution when working with sales or income data.

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.

Perfectly fine to not remove any data points here and good justification for not doing so. Actually one cool thing about unsupervised learning is that we could actually run our future analysis with these data points removed and with these data points included and see how the results change.

However based on your comment of

> "Yes, there are points considered outliers for more than one feature as 65, 75 and 154"

As you have discovered three of the data points that are considered to be outliers for more than one feature. But there are two more to find. There are 5 in total.

Optional: Could also look into doing this programmatically with the use of the Counter method

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.

Nice work with the cumulative explained variance for two and four dimensions. Could look into using `np.cumsum(pca.explainedvariance_ratio)`.

- As with two dimension we can easily visualize the data(as we do later)
- And with four components we retain much more information(great for new features)

And you have the right ideas in terms of the interpretation of these PCA components. I would recommend only mentioning the **most** prevalent features in each component(highest absolute magnitude). Therefore to go even further here:

- In terms of customers spending, since PCA deals with the variance of the data and the correlation between features, the first component would represent that we have some customers who purchase a lot of Milk, Grocery and Detergents_Paper products while other customers purchase very few amounts of Milk, Grocery and Detergents_Paper, hence spread in the data. So maybe this component could represent retail spending.

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.

Good comparison and good choice in GMM, as I would choose the same. As we can actually measure the level of uncertainty of our predictions! Would recommend mentioning the speed advantage for K-Means. As the main two differences in these two algorithms are the speed and structural information of each:

Speed:

- K-Mean much faster and much more scalable
- GMM slower since it has to incorporate information about the distributions of the data, thus it has to deal with the co-variance, mean, variance, and prior probabilities of the data, and also has to assign probabilities to belonging to each clusters.

Structure:

- K-Means straight boundaries (hard clustering)
- GMM you get much more structural information, thus you can measure how wide each cluster is, since it works on probabilities (soft clustering)

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.

Nice work. As we can clearly see that K = 2 gives the highest silhouette score. Would recommend doing this programmatically with the use of a for loop

```
for k in range(2,14):
   clusterer = GaussianMixture(n_components=k)
```

```
clusterer.fit(reduced_data)
....
```

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 am really sorry for having to mark this as *Requires Changes*, but it seems as the previous reviewer missed this, but make sure you also give some ideas in what **type of establishment** each cluster could represent. i.e. retailer, restaurant, etc...

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

Great justification for your predictions by comparing the purchasing behavior of the sample to the purchasing behavior of the cluster centroid!

Since you have used GMM, we can also check out the probabilities for belonging to each cluster

```
for i,j in enumerate(pca_samples):
    print "Probability of Sample {}: {}".format(i,clusterer.predict_proba([j])[0])
```

Conclusion

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

> "Then we would have to compare the results against those obtained from the original delivery service from the same segment. So in our case A would be the original delivery service applied to a segment S and B would be the new delivery service applied to the same segment S. "

This is key here! We should run separate A/B tests for each cluster independently. As if we were to use all of our customers we would essentially have multiple variables(different delivery methods and different purchasing behaviors).

https://en.wikipedia.org/wiki/A/B_testing#Segmentation_and_targeting
https://stats.stackexchange.com/questions/192752/clustering-and-a-b-testing

The wholesale distributor can then look at the p values for the tests that the null hypothesis that the difference between the chosen metric between the control group and the experiment is zero. If the p value for segment 0 A/B test is smaller, it means segment 0 customers are affected more by the change.

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

Nice idea to use the cluster assignment as new labels. Another cool idea would be to use a subset of the newly engineered PCA components as new features(great for curing the curse of dimensionality). PCA is really cool and seem almost like magic at time. Just wait till you work with hundreds of features and you can reduce them down into just a handful. This technique becomes very handy especially with images. There is actually a handwritten digits dataset, using the "famous MNIST data" where you do just this and can get around a 98% classification accuracy after doing so. This is a kaggle competition and if you want to learn more check it out here KAGGLE

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.

Would agree, real world data is really never perfectly linearly separable but it seems as our GMM algorithm did a decent job.

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Rate this review

Student FAQ