PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Meets Specifications

Very good analysis here and you do have a good understanding of these techniques!

If you wanted to go deeper, here might be some cool books to check out

- An Introduction to Statistical Learning Code is in R, but great for understanding
- elements of statistical learning More mathy
- Python Machine Learning I have this one, great intuitive ideas and goes through everything in code.

Based on your submitted comment of Interpretation of ICA components is based on the magnitude and sign of the components that my notebook produced. In last review, I could see that a component was referenced that was not part of my output so I couldn't quite follow that. The reference sources related to ICA were helpful and I guess I have interpreted the results properly this time. Any other pointers to improve my understanding will be much appreciated.

- You have done a great job here with your ICA interpretation. Just the 4th vector interpretation can be a bit confusing at times with the anti-correlation, but the - Cocktail Problem, somewhat shows this. ICA components are unique up to negation in some sense. So an ICA component that is +2A and -3B is the same as an ICA component that is -2A and +3B. ICA components really measure the linear differences between the original features. So, for example, a proper analysis would be more like "a bit of A vs B" instead of "some A but less B". Analyzing ICA components can be tricky, but you have done a good job.

Functionality

Code in required implementation sections produce the correct results. Results do not conflict with the answers provided.

Responses to Project Questions

At least one idea for what patterns might arise as components in PCA and ICA has been written.

Great job with both PCA and ICA as the first corresponds mostly to Fresh and the second Grocery. Also note that PCA deals with correlation between variables that create these individual eigenvectors. Therefore this is why we see Milk and Detergents also in the second as they are highly correlated to Grocery. And with ICA, different types of customers of the distributor is a great idea of what we would get back!

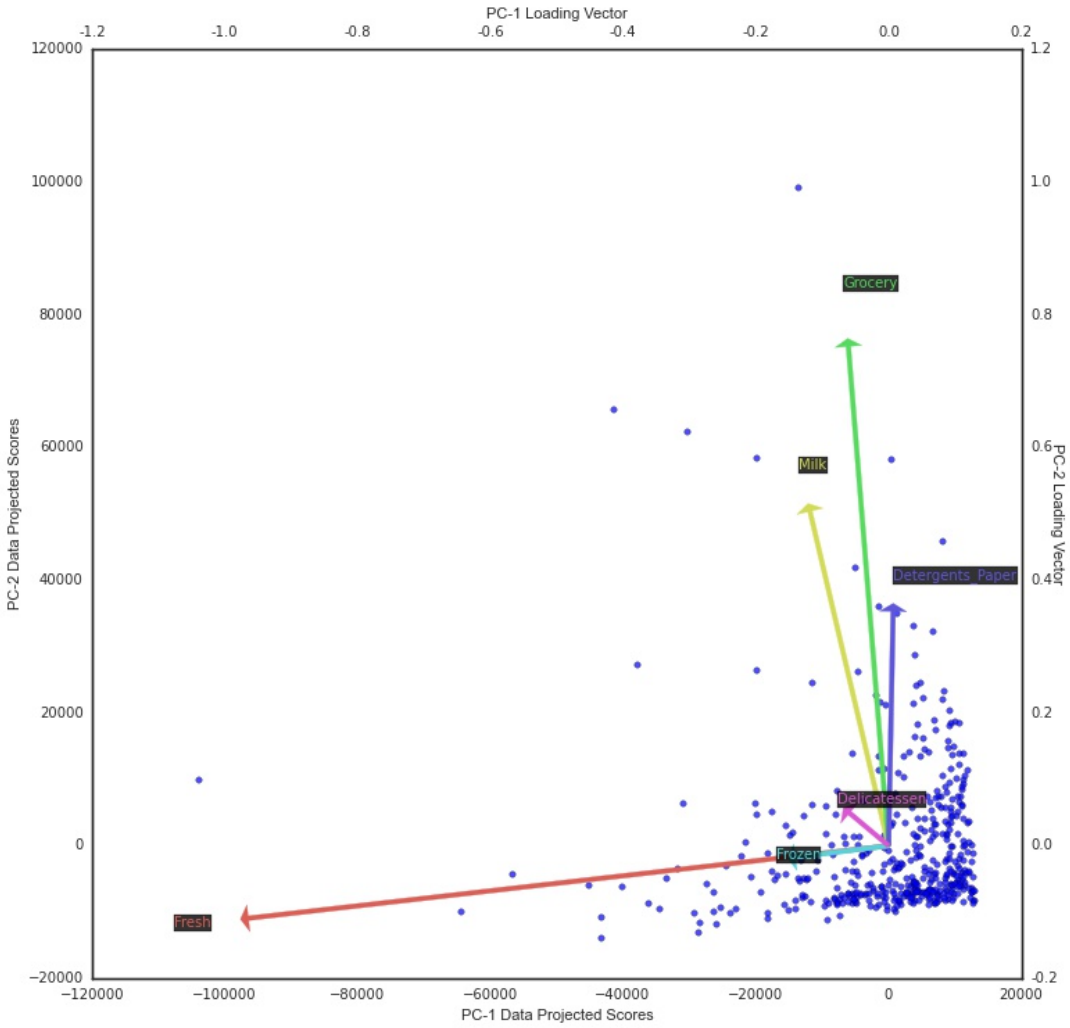
The variance explained by each dimension is appropriately reported and explained. A reasonable cutoff point for use has been explained.

Basis vectors for at least two PCA dimensions have been interpreted correctly and their significance has been discussed.

Nice interpretation of these vectors and what they really mean

Note: With the comment of " there seems to be some correlation among Frozen foods and Delicatessen i.e. customers buying Frozen food is likely to buy Delicatessen", as there is some correlation but not much. As this is more prevalent with the second dimension.

Note: With this statement of "Grocery is a favorite item among the high volume buyers unlike Fresh foods.", Since PCA works on variance, it wouldn't necessarily mean "a favorite item among the high volume buyers", but in terms of customers this first component really means that we have some customers who purchase a lot of Fresh products and others that purchase very little, hence spread of the data. PCA allows the business to determine the true underlying features that represent the data the best as display most influential features by representing them in the first and the second components. Therefore a company can really determine the best and even the worst predictors in the dataset. Thus consider your biplot and this one



Basis vectors for at least 4 dimensions of ICA have been interpreted correctly and their significance has been discussed.

Good work here with ICA and you have really captured the true essence of what these vectors represent.
I would say

- 1st - Frozen
- 2nd - Grocery
- 3rd - Del
- 4th - *Check this one out again* - Grocery with an inverse relationship with Deter_Paper. Therefore with this dimension with a large positive value for grocery and negative value for detergents, it would suggest that, independent of other effects, there is an anti-correlation between grocery and detergent purchases. Possibly this could come from a distinction between something like grocery stores and pharmacies which carry some food but far more paper products. Thus this means that these customers would purchase either Grocery or Deter but not both, hence inversely related!

Code Note: You can also easily visualize the linear differences between the features here with a bar plot

```
pd.DataFrame(ica.components_, columns=data.columns).plot(kind = 'bar', figsize = (12, 6))
```

Gaussian Mixtures and K Means have been compared. Student makes a choice which is justified based on the characteristics of the algorithms.

Great comparison. 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.

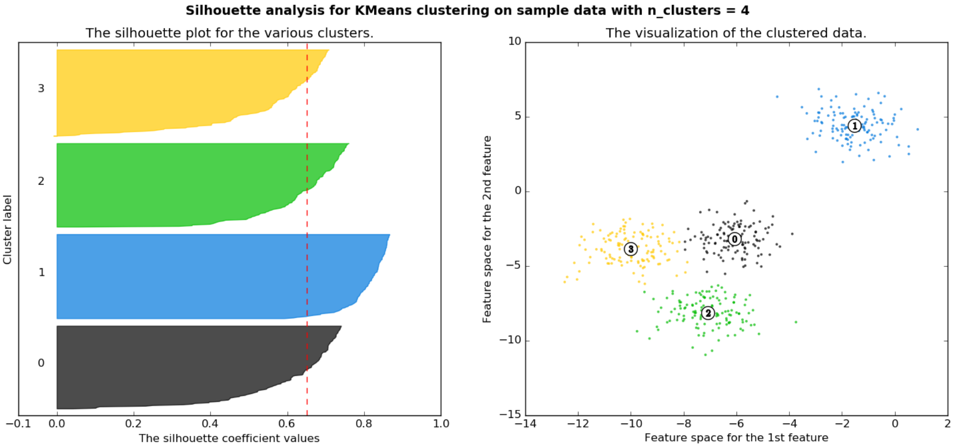
Structure:

- 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)

More than one choice of number of clusters has been tried out. Elements from each cluster have been sampled and interpreted.

Great work here with the elbow method to determine an appropriate cluster number and the pca.inverse_transform to to take the centroids and use the PCA model to bring the centroids into the original dimensions. I'm sure that this distributor would be happy to receive this analysis.

Additional Note: You can also determine an appropriate K with a silhouette score and you can produce a cool visual like this



Additional Note 2 : It would also be cool to reference the plot with the pca.inverse_transform analysis.

PCA has been used to visualize the data in two dimensions. A plot has been created which clearly shows different clusters. If clusters are not clearly visible some discussion has been made of how to improve the visualization.

Note here that one other way to improve this visual is to use a data transformation. As you can see that the data is very jumbled in the bottom right of the graph, if we were to plot the normalized data or a log transformation, we would essentially 'spread' the data out more.

Since our data is all highly skewed right, a log transformation could be very suitable, try clustering on this

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

One method has been discussed in detail and its usefulness has been discussed.

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

Conclusion

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

Great job implementing an A/B test here and I really like the example!

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

Nice, as we can use our cluster assignments as 'labels' and be able to run a supervised learning algorithm for new customers and place them into our clusters and be able to predict what they need.

DOWNLOAD PROJECT

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Student FAQ