UDACITY

Share your accomplishment!

Creating Customer Segments

Student Notes Project Review Project History

Does Not Meet Specifications

Functionality

SPECIFICATION
All code executes successfully and no errors are produced.

MEETS SPECIFICATION

Responses to Project Questions

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

MEETS SPECIFICATION

Reviewer Comments
Love the boxplots here as this is a great idea to see the spread of the data. With PCA, Fresh would definitely represent the first component as it has the highest variance in the data. Remember that PCA also deals with the correlation of variables, so you could also look at correlation matrix. For ICA, different types of customers of the distributor is a great idea of what we would get back!

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

MEETS SPECIFICATION

SPECIFICATION
Basis vectors for at least two PCA dimensions have been interpreted correctly and their significance

Reviewer Comments

I think you understand how to interpret these vectors, just your wording may be a little confusing here. Remember that the dimensions in Principle Component Analysis represent the combinations of the original features, and they should be interpreted the same way here. As I would say that

- The vector `[-0.97653685 -0.12118407 -0.06154039 -0.15236462 0.00705417 -0.06810471]` contains mostly data on fresh foods, plus some milk and frozen foods.
- The second dimension deals with milk, groceries, and detergent_paper, since these three magnitudes are much bigger than the rest

Note: Relook at this statement of "Looking at the biplot, we can see that most customers buy Fresh foods more than anything".

- Consider this - As the 1st principal component increases as the "Fresh" variable decreases. This component can be viewed as a measure of how "Fresh" orders decline with high-volume customers. Because of the high correlation for "Fresh" (-0.97), this principal component is primarily a measure of the "Fresh" variable. Therefore, high-volume customers order much less "Fresh" items (in comparison to other items) than low-volume customers. Since this vector is in the opposite direction here, this analysis would be switched.

Therefore can you adjust your answer based on these ideas!

SPECIFICATION

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

DOES NOT MEET SPECIFICATION

Reviewer Comments

Your interpretation of these vectors may be a bit off here, as the second vector here would not necessarily represent "predominantly grocery followed by some Fresh foods and frozen items", as the value of `0.04970109` for Fresh is way above the rest. To make these more interpretable, consider plotting the components with something like a bar plot

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

Code Note: You should also pass in all 6 components into FastICA as well, as this is not like PCA and the results will change based on how many components you assign `n_components=data.shape[1]`

Interpretation: You are correct with what each ICA could represent here. ICA components are tricky. First off, these components can be interpreted as: as the absolute value of the elements of the unmixing matrix increases, the corresponding feature has a strong effect on that components(which you have addressed). Therefore the magnitude of the coefficient represents prevalence in that feature. But we also need to take into account for the positive and negative sign of the coefficient as they are considered anti-correlated to each other. Meaning that if we have a high positive coefficient and a very negative coefficient, it would be interpreted as one or another but not both, hence anti-associated to each other.

Here are some helpful links:

- Udacity Forum
- Independent Component Analysis
- Coctail Problem.
    - This will demonstrate the anti-correlation between features, as when one sound gets louder the other lower

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

MEETS SPECIFICATION

Reviewer Comments
Love the statement of "GMM is generally fast to learn mixture models but fails when the number of samples is really high", as this is exactly correct and may people don't capture this idea! The core answer here is that K Means will give you much faster computations, but GMM gives you significantly more structural information and nuance. Being able to measure how ‘wide’ each cluster is in a sense is also a big advantage of GMM!

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

DOES NOT MEET SPECIFICATION

Reviewer Comments
I absolutely love the K-Means BIC elbow analysis here!!!! This is really cool as this analysis is not implemented in sklearn, you could even try to contribute to their repo with this!

The only thing you have left to do here is to interpret your clusters in terms of features. As you should answer your comment of "These customer types were identified by how much of the important product groups they buy relative to the average quantities." This should be pretty straight forward since you have already computed the pca.inverse_transform(centroids) to take the centroids and use the PCA model to bring the centroids into the original dimensions. Therefore go cluster by cluster and explain what each represents in terms of features. For example, "this cluster buys a lot of Fresh and not much else".

Code Note: You can also even easier interpret your pca.inverse_transform(centroids) by plotting them with a bar plot

```
pd.DataFrame(pca.inverse_transform(centroids)).plot(kind = 'bar')
```

SPECIFICATION
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.

MEETS SPECIFICATION

SPECIFICATION

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

MEETS SPECIFICATION

Reviewer Comments

Note: Relook at this statement of "predominant type of customer - a most repeating customer group buys more Fresh foods". As the first component of PCA deals with the highest variance in the data and would really tells us that we have some customers who purchase a lot of Fresh and other who purchase very little Fresh, hence spread of the data!

SPECIFICATION

Some method of improving the ability to get good results from an A/B test has been proposed.

MEETS SPECIFICATION

Reviewer Comments

Great job implementing an A/B test here. Just remember that we can only have one experimental variable at a time, as if we ever have multiple we won't know for sure what the true factor would be in our test.

SPECIFICATION

Some techniques that could be used in a supervised learning analysis have been proposed.

MEETS SPECIFICATION

Reviewer Comments

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.

Additional Reviewer Comments

This is a good analysis, just have a couple of thing here to address. But by doing all of these things you will get a much better understanding of how these techniques work.

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.