#Machine Learning Engineer Nanodegree

##Project 3: Creating Customer Segments

###Introduction

This documents presents the results for the third project w ithin the Machine Learning Engineer Nanodegree program. This assessment required the student to apply dimensionality reduction and classification techniques to a set of sales data from a w holesale grocery distributor, in order to categorize its customers into segments.

###Data

```
Table 1: Raw Dataset (first 5 records)
```

Out[6]:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	12669	9656	7561	214	2674	1338
1	7057	9810	9568	1762	3293	1776
2	6353	8808	7684	2405	3516	7844
3	13265	1196	4221	6404	507	1788
4	22615	5410	7198	3915	1777	5185

```
Dataset rows: 440
Dataset columns: 6
```

This assessment uses data provided via the Udacity platform. The dataset has 440 row s and 6 columns.

###Question 1

Before doing any computations, w hat do you think w ill show up in your computations? List one or tw o ideas for w hat might show up as the first PCA dimensions, or w hat type of vectors w ill show up as ICA dimensions.

####Answ er

Under the PCA method, the first dimension w ill show the global direction of the data (maximum variance). According the dataset variance's show n below , it may be reasonable to suggest that the first component w ill carry a high load of the 'Fresh' feature since it has the highest variance.

```
Table 2: Raw Dataset Statistics
```

Out[8]:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
std dev	12647.329	7380.377	9503.163	4854.673	4767.854	2820.106

ICA instead finds subcomponents according to w hich are statistically independent. Under this premise, it may be reasonable to suggest that the ICA method may also find 'Fresh' to be one of its components, how ever it is difficult to say w hich remaining components it w ould select as the components need not be orthogonal in the feature space, as opposed to PCA.

###Question 2

How quickly does the variance drop off by dimension? If you were to use PCA on this dataset, how many dimensions would you choose for your analysis? Why?

####Answer

Fit PCA.

```
Out[9]: PCA(copy=True, n_components=6, whiten=False)
```

Show PCA components.

```
Table 3: PCA Components
```

Out[10]:

	fresh	milk	grocery	frozen	detergentspaper	delicatessen
0	-0.977	-0.121	-0.062	-0.152	0.007	-0.068
1	-0.111	0.516	0.765	-0.019	0.365	0.057
2	-0.179	0.510	-0.276	0.714	-0.204	0.283
3	-0.042	-0.646	0.375	0.646	0.149	-0.020
4	0.016	0.203	-0.160	0.220	0.208	-0.917
5	-0.016	0.033	0.411	-0.013	-0.871	-0.265

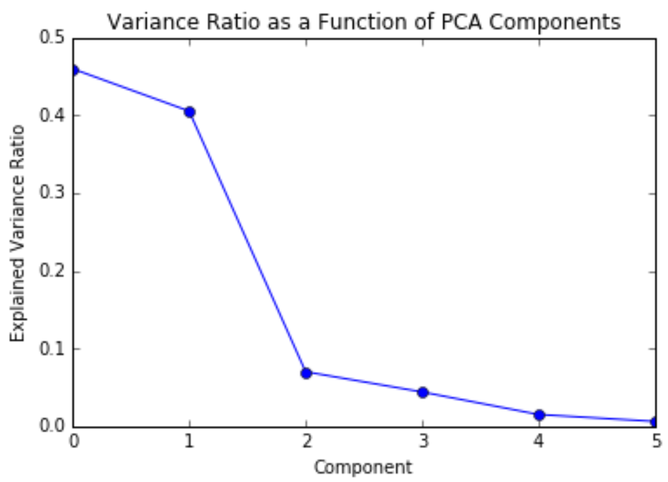
Calculate PCA dimensions.

```
Table 4: PCA Dimensions
```

Out[12]:

	dimension1	dimension2	dimension3	dimension4	dimension5	dimension6
0	0.46	0.405	0.07	0.044	0.015	0.006

Plot the expected variance ratio for each component.



The variance ratio of the first tw o dimensions are 0.4596 and 0.4052 respectively, resulting in 0.8648 of the variance being along the first tw o dimensions. The variance along the third dimension is only 0.0700, w hich is a 0.1728 drop relative to the second dimension. We should therefore keep only the first tw o dimensions for our analysis. This w ill help to avoid the curse of dimensionality.

Interpreting these results allow s us to form an initial impression about the customer segments contained in the data. One possibility could be that many customers are able to be split into customers ordering mainly 'Fresh' items versus customers ordering ordering 'Grocery', 'Milk' and 'Detergents'.

###Question 3

What do the dimensions seem to represent? How can you use this information?

####Answ er

These dimensions are linear combinations of the datasets five features. We can read off these linear combinations from the pca component array show n above.

Observing the first row , w e see that the first dimension is almost completely (anti-) aligned along the 'Fresh' direction (-0.9765) and slightly in the 'Milk' direction (-0.1212) and the 'Frozen' direction (-0.1524). It is almost orthogonal to the other directions.

Observing the second row , w e see that the second dimension is most strongly aligned in the 'Grocery' direction (0.7646) as w ell as the 'Milk' direction (0.5158) and 'Detergents_paper' direction (0.3654), and slightly in the 'Fresh' direction (-0.1106). It is almost orthogonal to the other directions.

###Question 4

For each vector in the ICA decomposition, w rite a sentence or tw o explaining w hat sort of object or property it corresponds to. What could these components be used for?

####Answ er

Fit ICA.

```
Out[15]: FastICA(algorithm='parallel', fun='logcosh', fun_args=None, max_iter=200,
             n_components=6, random_state=None, tol=0.0001, w_init=None,
             whiten=True)
```

Show ICA components.

Table 5: ICA Components (10^5)

Out[16]:

	fresh	milk	grocery	frozen	detergentspaper	delicatessen
0	-0.396	0.095	0.047	0.071	-0.160	0.103
1	0.030	-0.104	-1.365	0.132	2.778	0.538
2	-0.039	-0.031	-0.061	-0.053	0.044	1.818
3	-0.022	0.699	-0.527	-0.076	-0.460	-0.332
4	-0.034	-0.746	0.252	0.008	-0.401	0.536
5	-0.086	-0.019	0.087	1.115	-0.060	-0.594

Observing the first row, we see that the first vector has a high representation of the 'Fresh' feature, with a coefficient of -0.3965. The other features have weaker representation on the first dimension.

Observing the second row, we see that the second vector has a high representation of both 'Grocery' and 'Detergents paper', however the coefficients of each are of different polarity (-1.3651 and 2.7780 respectively). This indicates that, with all else held equal, high spending in 'Grocery' is associated with low spending in 'Detergents paper', and visa versa.

Observing the third row, we see that the third vector has a high representation of the 'Delicatessen' feature, with a negative coefficient of 1.8180. Indicating low spending on this feature within the vector.

Finally, observing the forth row, we see that the forth vector has a high representation of the 'Milk' feature, with a positive coefficient of 0.6986. Indicating high spending on this feature within the vector.

Like with the PCA analysis shown above, interpreting these results allows us to form an initial impression about the customer segments contained in the data.

###Question 5

What are the advantages of using K Means clustering or Gaussian Mixture Models?

####Answer

Gaussian Mixture Models (GMMs) make a probabilistic assignment of points to classes, whereas K Means makes a deterministic assignment.

An advantage of GMMs is that prior uncertainty about the assignment of a point to a cluster can be inherently reflected in the probabilistic model (soft assignment). Therefore, a GMM would be more suitable if the underlying dataset is a representation of a mixture of Gaussians. Alternatively, if cluster assignments are expected to be deterministic, a K Means clustering algorithm has advantages.

It is worth noting, when assessing computation speed, the EM algorithm with Gaussian mixtures is generally slightly slower than Lloyd's algorithm for K Means, since computing the normal probability (EM) is generally slower than computing the L2-norm (K Means).

Since there is nothing to suggest that this data has been formed from a mixture of normals, and the goal of this assessment is to discretely characterise customers (not to assign probabilities), K Means would be a more suitable algorithm to apply.

###Question 6

Generate a cluster visualization of the data.

####Answer

Fit PCA with two components.

```
Out[18]: PCA(copy=True, n_components=2, whiten=False)
```

Use PCA to transform the data under two components.

Table 6: PCA Reduced Data (first 5 records)

Out[19]:

	component1	component2
0	-650.022	1585.519
1	4426.805	4042.452
2	4841.999	2578.762
3	-990.346	-6279.806
4	-10657.999	-2159.726

Fit KMeans clustering algorithm w ith five clusters to the PCA reduced dataset.

Out[20]:
```
KMeans(copy_x=True, init='k-means++', max_iter=300, n_clusters=5, n_init=1
0,
    n_jobs=1, precompute_distances='auto', random_state=None, tol=0.0001,
    verbose=0)
```

Find the centroids for KMeans clusters.

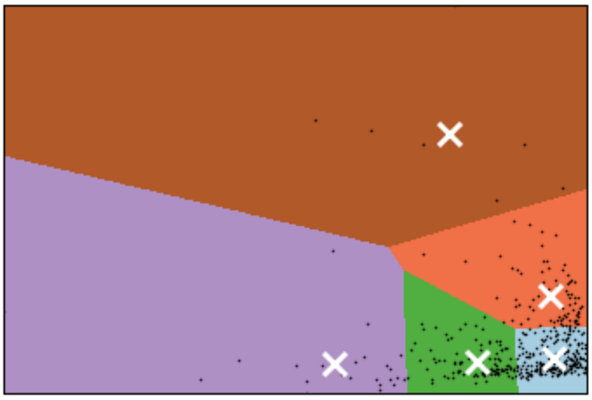
Table 7: Centoids for KMeans

Out[21]:

	component1	component2
0	6399.712	-4169.297
1	-9052.400	-4808.559
2	5607.917	14199.180
3	-37704.642	-5488.354
4	-14537.718	61715.671

Plot the KMeans clusters w ith centoids over the PCA reduced dataset.

Clustering on the wholesale grocery dataset (PCA-reduced data)
Centroids are marked with white cross



###Question 7

What are the central objects in each cluster? Describe them as customers.

####Answ er

Convert centroids back to original space.

Table 8: Converted Centroids

Out[23]:

	fresh	milk	grocery	frozen	detergentspaper	delicatessen	total
0	6211.925	2870.190	4369.565	2174.906	1403.382	851.040	17881.008
1	21372.193	4413.007	4831.711	4541.230	1060.825	1866.913	38085.879
2	4953.334	12440.644	18462.948	1951.626	8108.734	1953.423	47870.709
3	49427.357	7534.563	6075.205	8919.546	610.344	3779.463	76346.478
4	19370.306	39391.082	56034.130	4131.435	25326.809	6037.639	150291.401

These converted centoids can be used to represent five classified customer segments, along with their expenses under each feature.

Observing the first row, we see that the first customer spends the most on 'Fresh', followed by 'Grocery', 6211.9254 and 4369.5655 respectively. The same customer spends the least on 'Delicatessen', 851.0398. This customer has the lowest spend of all five customer segments.

Observing the second row, we see that the second customer spends the most on 'Fresh', followed by 'Grocery', 21372.1927 and 4831.7105 respectively. The same customer spends the least on 'Detergents Paper', 1060.8253. This customer has the second highest spend of all five customer segments.

Observing the third row, we see that the third customer spends the most on 'Grocery', followed by 'Milk'. Observing the fourth row, we see that the fourth customer spends the most on 'Frozen', followed by 'Fresh'. And finally, observing the fifth row, we see that the fifth customer spends the most on 'Milk' followed by 'Detergents Paper'.

###Question 8

Which of these techniques did you feel gave you the most insight into the data?

####Answer

The PCA data reduction technique in combination with K-means-clustering technique provided the most insight into the data. The PCA reduction technique was able to reduce the six features of the original dataset according to two components which maximize variance. This enabled the data to be clustered within a 2-dimensional space using a K-means-clustering technique.

By clustering the data and forming a number of customer segments, the client can now better disaggregate their customers by spending behaviour. Follow on analysis may attempt to draw relationships between spending behaviors of these customer segments. For example, of those customers who have a high total spend, which type of product do they spend the most on?

###Question 9

How would you use that technique to help the company design new experiments?

####Answer

If the client had plans to introduce or retire delivery strategies in the future, he or she could use the customer segmentation provided by the K-means technique in order to focus efforts on that specific segment. For example, the client could test whether a change in delivery strategy was able to shift item preferences for low spending customer segments from one particular item to another. The ability to understand preferences of each customer segment can help drive customer delivery experiences and ultimately improve customer experience.

###Question 10

How would you use that data to help you predict future customer needs?

####Answer

The client could for example, monitor changes in the spending behavior of each of these identified customer segments, and potentially introduce supervised learning techniques in order to predict changes in product demand. Doing so would provide the client with a competitive advantage in being able to lead changes (i.e. adjust inventories, investment and staff training) which suit shifting customer preferences.