1) In this section you will be using PCA and ICA to start to understand the structure of the data. Before doing any computations, what do you think will show up in your computations? List one or two ideas for what might show up as the first PCA dimensions, or what type of vectors will show up as ICA dimensions.

Answer:

PCA: PCA is use to find correlation in the data and it does that by finding the axis which maximizes the variance. If we take a look at the histogram presented above we can see that the first principle component will explain Fresh products because Fresh products have the maximum variance. The second principle component will explain Grocery and so on.

ICA: It tries to find a linear transformation of feature space into a new feature space such that each of the individual new features are statically independent. In this case ICA will show which commodities are independent of each other. The ICA algorithm will return a six by six matrix of all the features where each row tells us which features are independent of other in each component and this information can be used to divide the customers into different groups.

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: The variance drops of after the first two principle component. Since the first two principle components explain almost 86% of the data. It is good to take only the first two out of the six principle components.

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

Answer: The dimensions represents the principle components. The first row represents the first component. The second row represents the second component and so on.

We can see how the principle components explain the data and then select the once which explain most of the data. In this case the first two principle components explain 86% of the data.

The first principle component explains mostly(in descending order) Fresh, Frozen, Milk. The second principle component explains mostly(in descending order) Grocery, Milk, Detergent paper. Now that we know this we can use the first two PCA components that explain most of the data in the data set without much data loss. Therefore, PCA helps us in dimensionality reduction.

4) For each vector in the ICA decomposition, write a sentence or two explaining what sort of object or property it corresponds to. What could these components be used for?

Answer: ICA is used to separate individual features from a multivariate features. This allows us to define clear boundaries while analyzing the data. The first ICA contains Fresh and Detergent paper and it tells us that as Fresh increases Detergent paper also increases. The second ICA contains Grocery. The third contains Milk, Grocery, Delicatessen and as number of Milk increases Grocery and Delicatessen products decreases. The fourth contains Delicatessen. The fifth contains Detergent paper. The sixth contains Frozen and Delicatessen and as number of Frozen product decreases Delicatessen product increases. This information can be used to divide the customer base into different clusters.

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

Answer:

Advantages of K Means algorithm[1]:

1) It is fast, easy to understand and robust.

2) It is relatively efficient. It's efficiency is O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.

3) It gives best results when datasets are separated from each other.

4) It does hard assignment which means that it assigns data points into one class exclusively.

Advantages of Gaussian Mixture Model Classifier (GMM):

1) It does soft assignment which means that it gives probability of a data point belonging to a class instead of assigning a point to one class exclusively. This is good as there can be datapoint which can have features of various classes.

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

Answer: The central objects in each cluster represents average customer in that particular category. There are two main categories of customer are the small business and the large business. The cross on the right represents the small businesses and the one on the left represents the big businesses.

Number of clusters

After plotting several graphs of different number of clusters we take the plot with two clusters because as we increase the number of cluster to three, four or higher we can observe that the inter cluster distance is decreasing which seems like overfitting. Therefore, it is appropriate to take the plot with two clusters.

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

Answer: Principle component analysis was used to help in the reduction of dimensions. This was done by selecting the first two principle components which explained about 86% of the data. This helped is dimensionality reduction and helped us visualize the data. By reducing the dimensionality. After PCA analysis Gaussian mixture model was used to discover the different clusters. Unlike the K means algorithm, which does hard assignment, the gaussian mixture model does soft assignments which is good when the boundary between different clusters is not clear. Since in this data set the boundary between the customers are blurred it is good to use Gaussian mixture model instead of K means.

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

Answer: After separating the two classes of customers we can try changes in one class of customers without effecting the other class to optimize the needs of the customers. For example if we want to change delivery options they can change it for one class without affecting the other class.

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

Answer: Using unsupervised learning we have separated the customers into two different categories. Now we can label the customers in the two different categories and use this(customers in different category separately) as input to a supervised learning algorithm which can then be used to determine what kind of products or delivery method a future customer might prefer.

References:
[1]https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm