Customer Segmentation using K-means and Gaussian Mixture Model Clustering

Daniele Pennacchia & Rory Hurley

Introduction, Description and Motivation of the Problem

Customer segmentation analysis is hugely valuable in the retail industry. In such a dynamic environment a competitive enterprise should try to understand its customers by gaining insight into their needs, attitudes, and behaviours *(Tsiptsis and Chorianopoulos, 2009)*. Many retailers therefore benefit from the use of unsupervised machine learning techniques to segment similar customers into different groups.

The aim of this project is to critically evaluate two unsupervised machine learning algorithms, K-Means and Gaussian Mixture Models, to gain insight into customers in a typical online retail dataset. We will use appropriate performance measures and visual criteria to evaluate the models.

Data and Preprocessing:

- Dataset: UCI Online Retail. The structure consists of 541909 observations with 8 features.
- Dataset was highly unbalanced by country, so data pre-processing following similar strategy to Chen et al. *(2012)* reduced the observations to 3863 UK customers, with 3 features derived; 'Recency' (how recently a customer purchased), 'Frequency' (how often the customer purchased), 'Monetary' (total money spent by the customer).
- The 3 derived features were log normalised *(Figure 1)*.

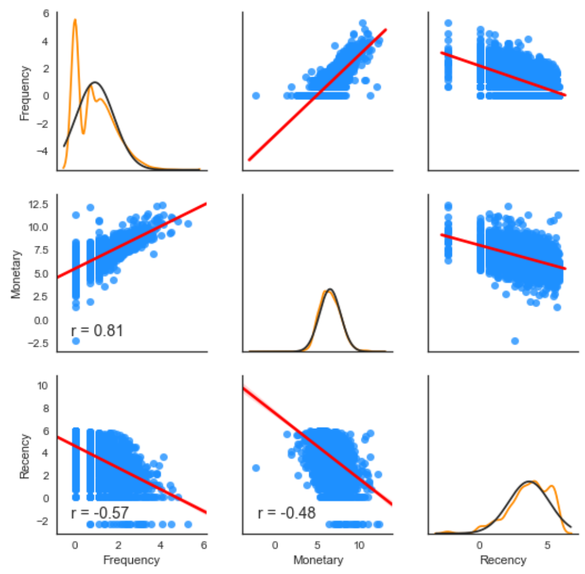

Figure 1: Recency, Frequency, Monetary Correlation matrix following log normalisation. Monetary is most normally distributed.

K-means

The k-means algorithm is one of the most commonly used algorithms for unsupervised clustering. It is partition-based clustering method where "the goal is to partition the data set into some number K of clusters" *(Bishop, 2006)*.

Advantages

- Simple and relatively easy to understand and implement
- Quickly labels clusters
- Generally converges in less iterations than other unsupervised algorithms

Disadvantages

- Non-probabilistic nature leads to poor performance for many real-world situations. It has no intrinsic measure of probability or uncertainty of cluster assignments (although it may be possible to use a bootstrap approach to estimate this uncertainty)
- No built-in way of accounting for oblong or elliptical clusters
- Sensitive to outliers due to the calculation of mean. Not good for clustering 'messy' data with irregular clusters *(Chen et al., 2012)*

Gaussian Mixture Models (GMMs)

GMMs cluster by assigning query data points to the multivariate normal components that maximize the component posterior probability given the data. GMMs are a form of soft partition based clustering, in which a score is assigned to each data point based on the association strength of the data point to the cluster *(Mathworks)*.

Advantages

- Probabilistic. Each data point is assigned a probability of it belonging to its cluster
- Fastest algorithm for learning mixture models
- Works well for non-spherically shaped clusters

Disadvantages

- Since it is Gaussian, it can assume that data is normally distributed an therefore may not perform well for other distributions and thus is sensitive to violations of distribution assumptions *(Hou., 2015)*
- When one has insufficiently many points per mixture, estimating the covariance matrices becomes difficult, and the algorithm is known to diverge and find solutions with infinite likelihood unless one regularizes the covariance artificially

K-means Methodology

- Use WCSS to find optimal number of K clusters
- Loop K-means algorithm with varying distance metrics ('sqeuclidean', 'cityblock', 'cosine', 'correlation') and 2 to 9 numbers of clusters.
- Algorithm hyperparameters set to 'kmeans++' for cluster centre initialisation and 'Replicates' set to 5.
- Evaluated using silhouette score, silhouette plot, % variance explained. 'evalclusters' used with silhouette criteria.
- From findings from analysis, give recommendation for optimal method, number of clusters and cluster metrics.

GMM Methodology

- Fitted GM distribution to data setting max iteration to 10000 (max number of iteration for convergence)
- Fitted GM model with k components from 2 to 9.
- Used hyper-parameters: Start = 'plus' which implements 'kmeans++' for cluster centre initialisation. Regularization value = 0.1 to avoid ill conditions covariance estimates and non-convergence. Covariance type 'full'
- Repeated 10 times to find average silhouette score per iteration.
- Evaluated using silhouette score, silhouette plot and visual analysis

Hypotheses

- Using domain knowledge, we believe that the more suitable algorithm will segment the customers into around 3-6 clusters in order to provide applicable insight to the industry.
- We expect more unclear partitioning with GMM due to the non-normal distribution of the features.
- For k-means, we expect squared euclidean ('sqeuclidean') distance hyper parameter to produce more robust clustering results because the data is log transformed, rather than normalised like cosine. Further, low correlation between features (figure 1) indicates that 'correlation' distance may perform poorly.

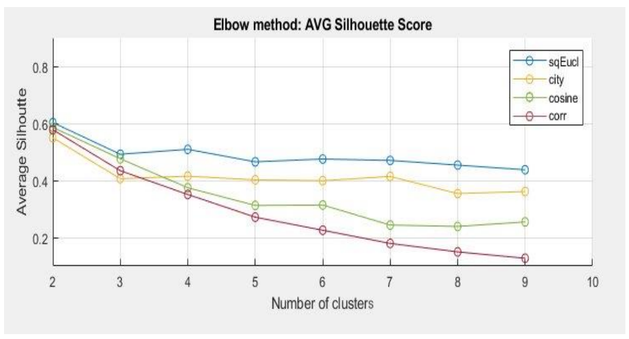

Figure 2: K-means average silhouette score for different distance parameters.

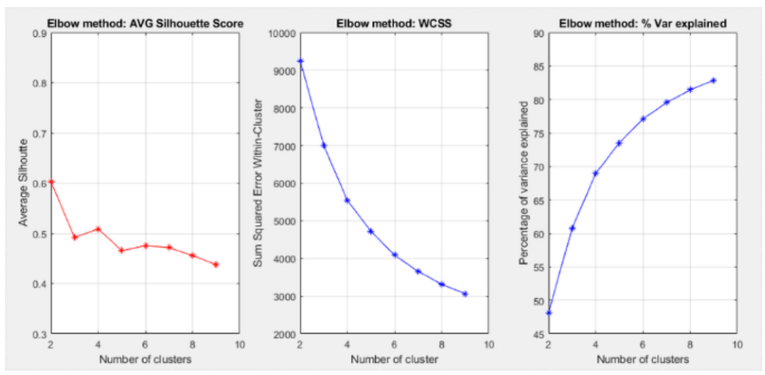

Figure 3: K-means elbow method for average silhouette score, WCSS, % variance explained

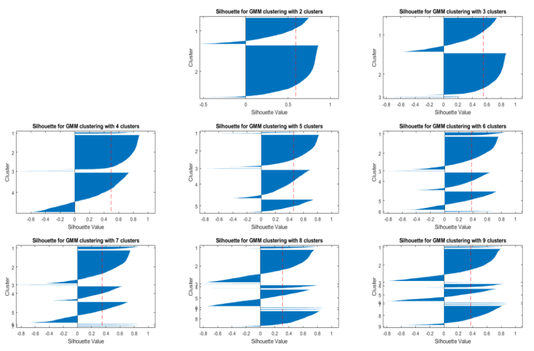

Figure 4: GMM silhouette score plot

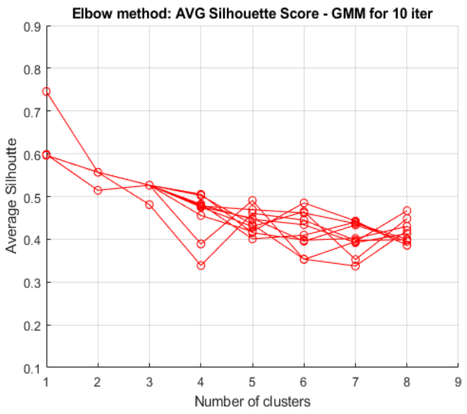

Figure 5: GMM average silhouette scores for 10 iterations.

Results & Evaluation

- The results show that the most suitable distance parameter was squared Euclidean ('sqeuclidean') for K-means. This is seen in Figure 2.
- Results shown in Figure 3 indicate that the optimum number of clusters for k-means is 4. By taking into account the average silhouette score, WCSS and average variance explained, we can observe a rise and fall when K=4, a significant reduction in gradient of the WCSS line when K=4 and a plateauing effect in the average variance explained.
- Similar tests were run using a GMM model. The results for these tests proved much more inconclusive compared to K-means. From Figure 4, there appears to be a lot more 'misclassification' and instability of silhouette scores as the number of clusters changes.
- As k increased throughout GMM iterations, the general observed trend was that silhouette score decreased. A high score was observed when K = 2.
- A possible reason for the poorer performance of GMM compared to K-means could be attributed to the non-normal distribution of the data (Figure 1).
- On reflection, the analytical approach taken by Chen et al. (2012) is valid, and likely to produce optimal results for customer segmentation.

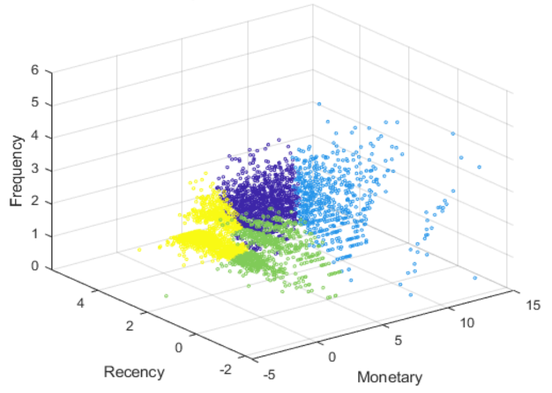

Figure 6: 3D RFM segmentation using optimum k-means parameters.

Conclusion & Future Work

- From our research, it is believed that the distribution plays a key role in unsupervised learning, especially when using probabilistic methods such as GMM.
- Future work on K-means would be to initialise cluster centroids using medoids (k-medoids), as these are less sensitive to outliers.
- Outlier detection during the pre-processing stage for both algorithms to observe effect on performance.
- Use an ensemble of the two methods to determine likelihood (using GMM) on the final customer clusters (from K-means), for precise marketing strategies.

References

1. Tsiptsis, K. and Chorianopoulos, A. (2009) *Data Mining Techniques in CRM: Inside Customer Segmentation*. A John Wiley and Sons, Ltd. ISBN: 978-0-470-74397-3.
2. Chen, D., Sain, S.L., Guo, K. (2012) 'Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining', *Journal of Database Marketing & Customer Strategy Management*, Volume 19, Issue 3, pp. 197-208.
3. Bishop, C. (2006) *Pattern Recognition and Machine Learning.* New York: Springer.
4. Mathworks (2018) *Gaussian Mixture Models.* Available at: https://uk.mathworks.com/help/stats/gaussian-mixture-models-2.html (Accessed: 23/11/2018).
5. Hou, W. (2015) *K-Means vs GMM & PLSA. San Jose State Univerisity. Available at:* http://www.math.sjsu.edu/~gchen/Math285F15/MATH285_Project_Report_Weiqian.pdf (Accessed: 23/11/2018)