Clustering: A beginner's guide to K-means and Gaussian Mixture Models

*

Omik Save
School of Energy Matter and Transport Engineering
Arizona State University
Tempe, United States
osave@asu.edu

Gautam Sharma
CIDSE
Arizona State University
Tempe, United States
gsharm17@asu.edu

Hrishi Patel
ECEE
Arizona State University
Tempe, United States
gsharm17@asu.edu

***Abstract*—Clustering is an important technique to segregate similar data into the same classes. Unsupervised clustering techniques like k-means are used to understand and model behaviour of humans and club them together with the hope that they get along. Such information can be very critical for hospitality and tourism sectors to boost their revenues. The problem addressed in this report is analyzing customers in a shopping mall and give recommendation to the marketing team such that the marketing strategies attract the most likely spenders. It was found that males and females in the age category (22-34) were most likely to buy money in the mall. The clustering was performed using k-means and Gaussian Mixture Model (GMM) with optimal clusters as 6 for each.**

***Index Terms*—Clustering, cluster analysis, K-means, GMM, EM, centroid, distribution.**

I. Data Preparation

The data was obtained from [1] and consists records of customers at the shopping mall of interest. 200 observations of various candidates were recorded that is distributed across 4 important features, namely, a) Gender b) Age c) Annual income (k$) d) Spending score. The 'spending score' feature is a metric assigned to each customer by the mall, based on their behaviour and spending nature.

While age, annual income and spending score is a numerical data and is easier to compute, gender on the other hand being categorical data needs to be addressed in terms of a number. Since gender is distributed across two classes only, i.e. male and female, it may seem lucrative to represent male by the number 1 and female by the number 0. However, converting the gender feature into a binary system would result in bias in the unsupervised clustering models. Since K-means compute distance and GMM finds probability, these models may certainly treat the binary system of gender feature unfairly. Although the goal is to represent gender in terms of number, it should be noted that no bias should be involved during the process. The solution to this problem is to treat each category as a feature. By implementing this every observation would be represented by two new features, i.e. male and female as opposed to the original representation of single gender feature. Thus, if a customer is male, the male feature would represent 1 and female feature would represent 0. Similarly, a female customer would be represented by 1 in female feature and 0 in male feature. Thus, all observations are now represented by 5 features, i.e. male, female, age, annual income and spending score. Figure 1 displays the improved encoding system for categorical data.

Gender		
26	Male	1
27	Female	0
28	Male	1

a.

Gender			
Obs. no.	Male	Female	
26	1	0	Male
27	0	1	Female

b.

Fig. 1. a) Binary encoding system imposes numerical bias. b) Categorical feature encoding system does not numerically alter categories.

It is also important to normalize the data in the age, annual income and spending score features to avoid any bias due to differences of values and also to obtain a homogeneous scale for the given data. Thus, a scale of 0 to 1 is chosen to normalize the features mentioned such that 0 is assigned to the lowest and 1 is assigned to the highest observation in each feature. Upon normalization, 5 features of observations are reduced to 2 features using principal component analysis (PCA) where only the first two principal components are chosen. Thus, the data is prepared for implementation in clustering models.

II. Algorithm Implementation

A. K-means Clustering

Initialization of k-means algorithm is extremely important for the quality of results. In this project we have used K-means++, a smart initialization technique to probabilistically

assure that the centroids are far from each other. Usually K-means++ takes more time than K-means but gives us better results. The major drwabcak of K-means is that it is extremely sensitive to the initialization of centres.

Algorithm 1: K-means++ clustering

Initialization: Randomly select the first centroid c_k
c_k is the last initialized centroid
prev=c_k
for $j = 1$ *to* $m_{observations}$ **do**
  $||x_j - c_k||^2$;
end
centroids.$append(c_k)$ Select the next centroid c_{k+1} s.t.:
for $j = 1$ *to* $m_{observations}$ **do**
  c_{k+1}=max:$||x_j - prev||^2$;
end
Repeat the above two steps until all the centroids have been initialized
K-means
for $i = 1$ *to* $m_{iterations}$ **do**
  for $j = 1$ *to* ç **do**
    $S_i^k = \text{min:}||x_j - c_k||^2$;
    $w_i^k = \frac{1}{|S_i^k|} \sum j x_j$
  end
  check; converge == False;;
  else; break;
end

B. Gaussian mixture model with expectation maximization

The following algorithm was used for our EM algorithm. Instead of randomly initializing the weights, we initialize the data by K-means and after getting the rough estimate we use EM algorithm on it. This step is necessary as since EM is a soft clustering algorithm, it is highly affected by the initialization.

Algorithm 2: EM algorithm

First we initialize the centroids using k means for a fixed number of iterations.
for $i = 1$ *to* $m_{iterations}$ **do**
  for $j = 1$ *to* $m_{observations}$ **do**
    $S_i^k = \text{min:}||x_j - c_k||^2$;
    $w_i^k = \frac{1}{|S_i^k|} \sum j x_j$
  end
end
Now we make use of EM algorithm
for $i = 1$ *to* $m_{iterations}$ **do**
  for $i = x_i$ *to* x_n **do**
    E step : $r_{ic} = \frac{\pi_c \mathcal{N}(x_i; \mu_c, \sum_c)}{\sum_{c'} \pi_{c'} \mathcal{N}(x_i; \mu_{c'}, \sum_{c'})}$;
    M step : $m_c = \sum_i r_{ic}$;
    $\pi_c = m_c/m$;
    $\Sigma_c = \left(\frac{1}{m_c}\right) \sum_{i=1}^{N} r_{ic} \cdot (x_i - \mu_c)^T (x_i - \mu_c)$
  end
  check; converge == False;;
  else; break;
end

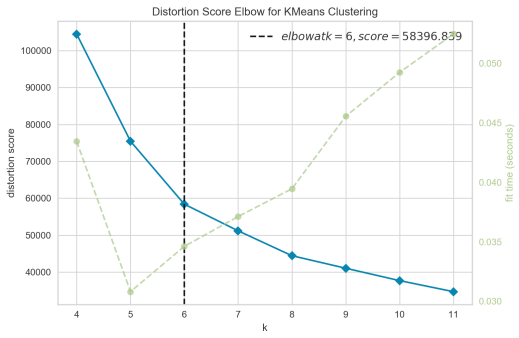

Fig. 2. Statistics of K-means cluster

III. Results

A. Model Performance

Before implementing the clustering model, it is important to understand the performance of model with changing parameters like number of clusters. Thus, optimum clustering has to be found in order to classify customers accurately. Various methods are used to calculate model performance. We chose the elbow method for K-means and log-likelihood method for GMM models.

Firstly, the elbow method is used to evaluate the K-means model as shown in fig 2 where y-axis represents sum of squared distances of all points from their assigned centers. This helps us visualize that increasing the number of clusters decrease the sum of squared distance. To find the optimum number of clusters, the point of elbow is chosen as the change in sum of squared distance from that point is not prominent. Hence, for this data, K-means model with 6 clusters was chosen.

For visualization purposes, plots of k-mean clustering with 4,6,8 and 10 (figure 3 - 6) clusters were also plotted as shown. The scatter plots represent principal component 1 on the x axis and principal component 2 on the y-axis. The legends for the scatter plot are as follows, cluster 1: red, cluster 2: green, cluster 3: blue, cluster 4: cyan, cluster 5: magenta, cluster 6: gold, cluster 7: lime, cluster 8: grey, cluster 9: orange, cluster 10: orchid(purple):

However based on the recommendation of elbow-plot, k-means with 6 cluster model was chosen as the optimum model. Since GMM takes probability into account, elbow plot is not justified to analyze optimality of GMM. We chose log-likelihood plots to measure performance of GMM clustering with increasing number of clusters. It should be noted that log-likelihood increases with increasing number of clusters. Log-likelihood defines the log of probability of the data x observed in a model $g[x]$. Thus, summation of all such logs gives us

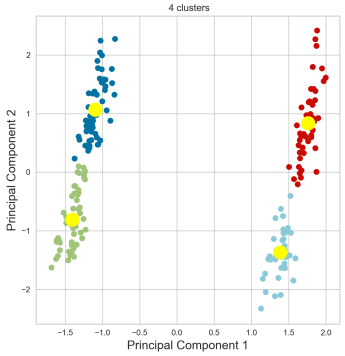

Fig. 3. K-means scatter plot with 4 clusters

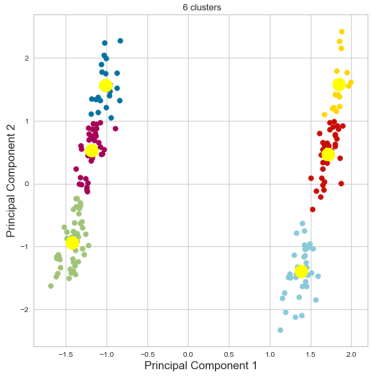

Fig. 4. K-means scatter plot with 6 clusters

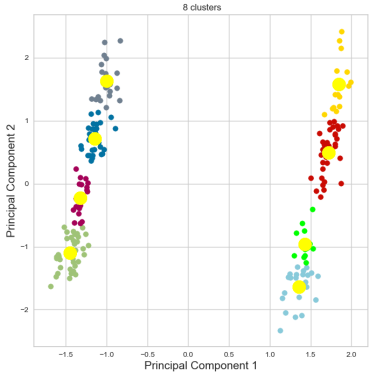

Fig. 5. K-means scatter plot with 8 clusters

the log-likelihood function for each clusters. As seen in figure 7, the slope increases rapidly from 4 clusters to 6 clusters. From 6 - 10 clusters, the slope does not increase with higher magnitude. Hence, 6 clusters is chosen to be optimum model since the slope increase tends to lessen beyond 6 in the log likelihood plot.

For visualization purposes, plots of GMM clustering with 4,6,8

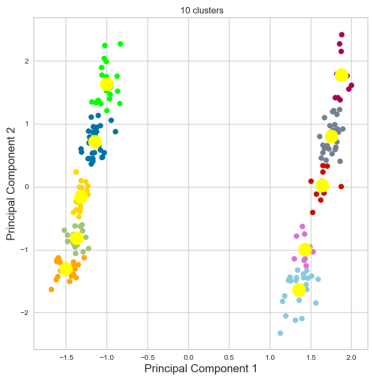

Fig. 6. K-means scatter plot with 10 clusters

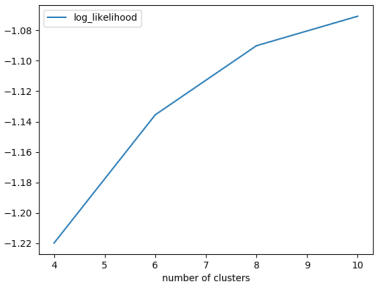

Fig. 7. Log-likelihood function for GMM clusters

and 10 (figure 8- 11) clusters were also plotted as shown. The scatter plots represent principal component 1 on the x axis and principal component 2 on the y-axis. The legends for the scatter plot are as follows, cluster 1: red, cluster 2: green, cluster 3: blue, cluster 4: cyan, cluster 5: magenta, cluster 6: gold, cluster 7: lime, cluster 8: grey, cluster 9: orange, cluster 10: orchid(purple):

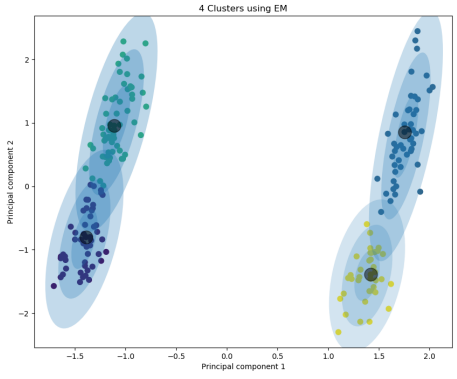

Fig. 8. GMM scatter plot with 4 clusters.

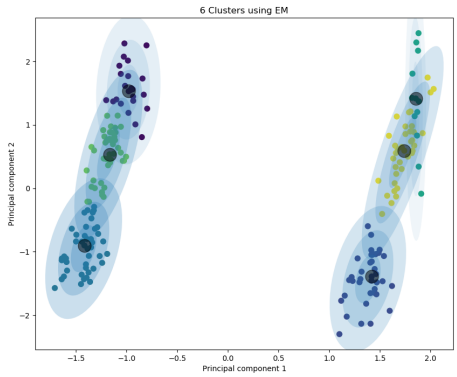

Fig. 9. GMM scatter plot with 6 clusters

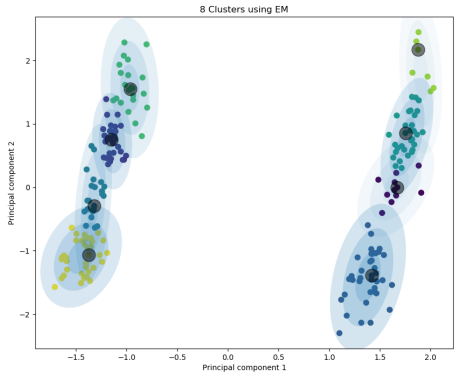

Fig. 10. GMM scatter plot with 8 clusters

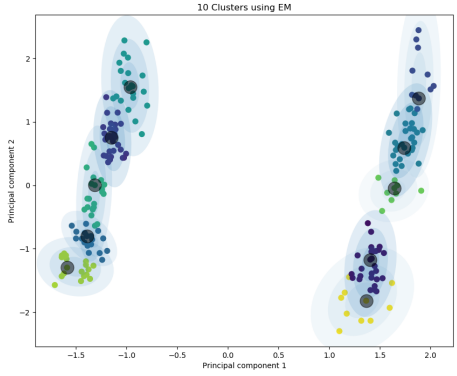

Fig. 11. GMM scatter plot with 10 clusters

B. Interpretation

It is evident from the model performance above that the optimum number of clusters for both K-means and GMM is 6 clusters. Interpreting the results require careful analysis of trend in each cluster. This section elucidates the clustering information in terms of features of the data provided.

1) K-means: The performance of K-means with 6 clusters was as follow:

	C1	C2	C3	C4	C5	C6
Gender Split (M/F)	0%/100%	100%/0%	0%/100%	100%/0%	0%/100%	100%/0%
Avg. Age (years)	53.4	43.588	28.125	28	40.513	61.933
Avg. Annual Income (k$)	57.6	67.176	60.6	62.2	58.641	51.066
Avg. Spending Score	26.88	31.323	**72.583**	**72.102**	41.41	26.133

Fig. 12. Statistics of K-means cluster

As seen in fig 12, the customers likely to shop from the mall lie in Cluster 3 and Cluster 4. The cluster 3 represents female customers with an average age of 28 years and an average salary of $ 60.6k. Similarly, Cluster 4 represents male customers with an average age of 28 years and average salary of $ 62.2k. It was also observed that candidates in these clusters below average age and salary were also most likely to shop. It was also observed that the standard deviation in annual income was very large (+/- 27k) while standard deviation in age was small (+/- 5 years) for these clusters. Age is hence a very important factor to target customers. Thus, based on the k-means model, the shopping mall should target males (21-35 years) and females (22-34 years) since they are the most likely to make purchases. It can also been seen from table that in cluster 2 where average age is high the spending score is low even though the average annual income is high. Thus, age influences tendency to make purchase more than annual income.

2) Gaussian Mixture Model: The performance of GMM with 6 clusters was as follows:

	C1	C2	C3	C4	C5	C6
Gender Split (M/F) (Total)	33.33/67.7% (24)	37.8/62.16% (37)	45.65/54.3% (46)	47.62/52.3% (42)	41.67/58.3% (12)	50/50% (38)
Avg. Age (years)	30.833	36.02	40.93	36	36.75	47.2
Avg. Annual Income (k$)	60.66	62	61.32	54.5714	58.4167	63.44
Avg. Spending Score	**72.458**	**67.027**	26.52	60.9524	**69.25**	29.657

Fig. 13. Statistics of GMM cluster

It is clear from the statistics that the clusters to most likely spend are Cluster1, Cluster2 and Cluster5 as seen in figure 13. It is also evident that female candidates are more likely to spend money than male candidates. The information gathered from dominant cluster i.e. 1 are that the target audience is dominantly females but males as well belonging to the age group 21 - 39 (calculated as +/- 8.9 years standard deviation from mean of 30.8 years) and average income in the range of $30 - 90 k (calculated as +/- $ 30k standard deviation from mean of $ 60.66 k) annually. As seen, the average annual income varies largely in the cluster and hence focus should more on the age factor. It can also be seen from the table that as average age increases in cluster 3, the spending score is

low even when the average annual income is high. Thus, age drives spending tendency more than the annual income.
The results are synonymous with K-means interpretation where the idea was to target males and females belonging to the age group (22-34 years).

IV. Recommendation

Finally, based on the two models of clustering the recommendations for the marketing team is as follows:

1) **K-means recommendation:** Design marketing strategies that attract customers male or female between the age group 21-35 years. The annual income may vary between $40 - 90 k. However, age directs the purchasing tendency more than the annual income of an individual.
2) **GMM recommendation:** Design marketing strategies to attract customers in the age group 21-39 years. Females are more dominantly to make a purchase than male customers. The annual income may vary between $ 30-90 k. Hence, age is a directs purchasing tendency of an individual than their average annual income.

Hence, the mall should advertise such that the young crowd is attracted.

References

[1] Vijay Choudhary. *Mall Customer Segmentation Data*. 2018. URL: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python#Mall_Customers.csv.