WEEK 8 ASSIGNMENT

Concepts of Statistics 2 – DATA-51200 | Spring 2 2020

Christina Morgenstern

---

1. On your own words, summarize (in less than one page) the steps of K-means clustering. Make sure to give example(s). What are the advantages and disadvantages of the K-means clustering? Any limitations?

The k-means algorithm is a popular nonhierarchical clustering algorithm that segments metric data into predefined number of *k* clusters. In the first step, the initialization, the algorithm chooses *k* random data points as initial centroids, with a centroid being the centre of a cluster. During the second step, cluster assignment, each data point is assigned to a cluster based on its distance to the centroid. The data point with the smallest distance to a centroid and thus closest to this cluster will be assigned to this cluster. Most often, the distance measure Euclidean distance is used to calculate the similarity within clusters. In step three, the clusters need their centroids to be updated. Therefore, the mean of all examples in a cluster will be the value of the new centroid. Steps two and three in this process are repeated until the centroids don´t change anymore and the data points have been separated into the *k* classes. Figure 1 shows a graphical representation of the k-means clustering process.

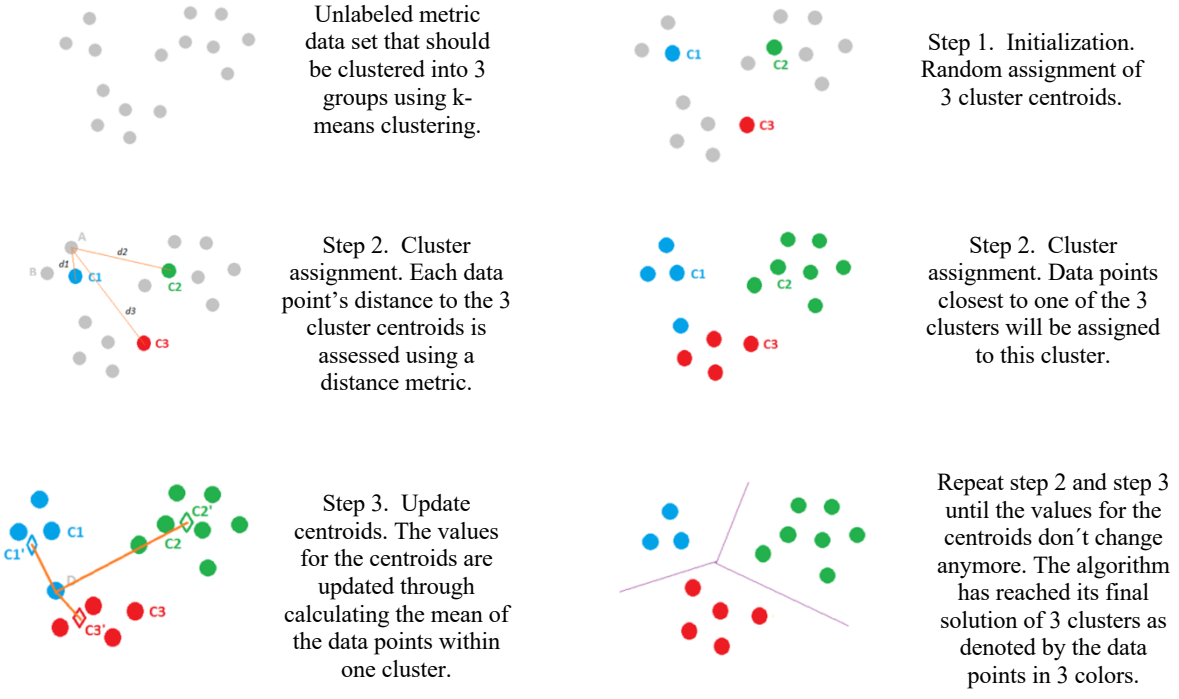

Figure 1. Visualization of the k-means clustering algorithm (pictures taken from: https://healthcare.ai/step-step-k-means-clustering/).

One major advantage of the k-means clustering is its application to large data sets because of the sequential processing and thus the avoidance of similarity matrices among all observations. Furthermore, results are less prone to outliers, the distance metric used or the inclusion of irrelevant variables. On the other hand, k-means can only realize its full potential when nonrandom seed points are used. Also, the clustering solutions do not guarantee an optimal solution and require further validation and analysis. The resulting clusters which most often are equally sized and spherical in shape are sometimes not ideal. However, the major limitation of k-means clustering is its application to metric data only.

References:
[1] Multivariate Data Analysis by Joseph F. Hair Jr, William C. Black, Barry J. Babin and Rolph E. Anderson, Pearson, 8th edition, 2019
https://towardsdatascience.com/clustering-using-k-means-algorithm-81da00f156f6 accessed on 8th of May 2020
https://healthcare.ai/step-step-k-means-clustering/ accessed on 8th of May 2020

2. For the data set associated with this homework (HBAT), apply K-means clustering (use Euclidean distance measure) to cluster the observations in the variables X6, X8, X12, X15 and X18. Report the results in case of 3 clusters, 4 clusters and 5 clusters. The results should include the Means of the Clusters, the Clusters Standard Deviations and Distance Between Cluster Centroids. In each case, report how many observations each cluster has. Which number of clusters (3, 4 or 5) gives the best results? Explain. (You may use any software and programming language you feel comfortable dealing with. Make sure to include your codes, diagrams and results)

The goal of this clustering assignment is to segment HBAT customers into groups based on four variables X6 (product quality), X8 (technical support), X12 (salesforce image), X15 (new product development) and X18 (delivery speed).

Using SAS Studio und the Tasks and Utilities function *Cluster Observations* on the HBAT data seta and the selected variables X6, X8, X12, X15 and X18, the following code was generated:

```
ods noproctitle;

/*** Standardize variables ***/
proc distance data=WORK.KMEANS stdonly outsdz=Work._Temp_sdz;
      var ratio(x6 x8 x12 x15 x18 / std=std);
run;

proc fastclus data=Work._Temp_sdz maxclusters=3 drift list distance;
      var x6 x8 x12 x15 x18;
run;

proc delete data=Work._Temp_sdz;
run;
```

The cluster analysis uses standard deviation as standardization method and applies Euclidean distance as dissimilarity measure. k-means clustering was used as clustering method with 3, 4 and 5 cluster solutions.

To cluster the data into 4 and 5 clusters, the same algorithm was run but using `maxclusters=4` and `maxclusters=5`, respectively.

The results of clustering using 3 clusters is summarized in Table 1 and shows that the frequency of observations is highest for cluster 3 (64 data points), followed with 24 observations for cluster 1 and 12 observations for cluster 2.

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	24	0.8472	2.8086		3	2.1083
2	12	0.7922	2.5150		3	2.3499
3	64	0.8882	3.4069		1	2.1083

Table 1. Summary of k-means clustering with 3 clusters on HBAT customers using variables X6, X8, X12, X15 and X18.

Cluster Means					
Cluster	x6	x8	x12	x15	x18
1	-0.114590252	0.088208959	1.214190050	0.142326315	0.904093656
2	-0.233955098	-1.322045385	-0.153561158	1.400937451	-0.128443024
3	0.086837925	0.214805150	-0.426528552	-0.316048140	-0.314952054

Cluster Standard Deviations					
Cluster	x6	x8	x12	x15	x18
1	1.274001090	0.736722094	0.743126157	0.675226959	0.643669578
2	1.002587193	0.672403577	0.649096092	0.766033476	0.819955136
3	0.883334424	0.954768423	0.738461756	0.905260584	0.942064441

Distance Between Cluster Centroids			
Nearest Cluster	1	2	3
1	.	2.554220554	2.108254496
2	2.554220554	.	2.349926396
3	2.108254496	2.349926396	.

Table 2. Cluster means, cluster standard deviations and distance between cluster centroids of k-means clustering on HBAT customers using variables X6, X8, X12, X15 and X18 and 3 clusters.

Repeating the k-means clustering with a 4-cluster setting shows the highest number of observations for cluster 2 (52), followed by cluster 4 with 25 observations, cluster 3 with 12 observations and cluster 1 with 11 observations (see Table 3 for cluster summary).

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	11	0.8522	2.7129		2	2.1939
2	52	0.8489	2.6640		4	1.8278
3	12	0.7853	2.6106		2	2.2357
4	25	0.8040	2.7770		2	1.8278

Table 3. Summary of k-means clustering with 4 clusters on HBAT customers using variables X6, X8, X12, X15 and X18.

Cluster Means					
Cluster	x6	x8	x12	x15	x18
1	0.240248881	0.201068906	1.903011036	0.520594649	0.600831134
2	0.137453211	0.074387042	-0.175679513	-0.154562514	0.500854479
3	-0.335415217	-1.365605365	-0.176875100	1.322797514	-0.219215480
4	-0.230612882	0.412295208	-0.387011422	-0.542514423	-1.200919585

Cluster Standard Deviations					
Cluster	x6	x8	x12	x15	x18
1	1.260315192	0.926395845	0.497452378	0.619739633	0.743508472
2	1.006318275	0.896395030	0.776923168	0.888159252	0.628000121
3	0.925499262	0.629884824	0.659817359	0.856357217	0.813074753
4	0.860003627	0.859098441	0.800823977	0.771448926	0.719926295

Distance Between Cluster Centroids				
Nearest Cluster	1	2	3	4
1	.	2.193946711	2.903065560	3.144365875
2	2.193946711	.	2.235682783	1.827801247
3	2.903065560	2.235682783	.	2.767526481
4	3.144365875	1.827801247	2.767526481	.

Table 4. Cluster means, cluster standard deviations and distance between cluster centroids of k-means clustering on HBAT variables X6, X8, X12, X15, X18 using 4 clusters.

The results of k-means clustering on HBAT customers using the chosen variables and 5 clusters shows that clusters 2, 3 and 4 have the highest number of observations with 34, 28 and 22 customers, respectively. This is followed by two small clusters 1 and 5 with 9 and 7 observations, respectively (see Table 5 for cluster summary).

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	9	0.6275	1.7250		3	2.2054
2	34	0.7692	2.8238		3	1.8877
3	28	0.8049	2.6110		2	1.8877
4	22	0.7797	2.8272		2	2.1745
5	7	0.7067	1.9627		3	2.6856

Table 5. Summary of k-means clustering with 5 clusters on HBAT customers using variables X6, X8, X12, X15 and X18.

Cluster Means

Cluster	x6	x8	x12	x15	x18
1	-1.542989574	0.639968702	0.880023562	0.167442723	1.108331680
2	0.702707869	0.461031138	-0.712638608	0.196991439	0.155220899
3	0.325354466	-0.327166562	0.924431069	0.035880584	0.461577936
4	-0.645221249	-0.021680990	-0.377516289	-0.849391269	-1.057829187
5	-0.702888421	-1.685304502	-0.181315850	1.353894019	-0.700633681

Cluster Standard Deviations

Cluster	x6	x8	x12	x15	x18
1	0.175835398	0.688485427	1.061232569	0.475961647	0.333519298
2	0.553739022	0.804042840	0.647772249	1.013103399	0.748028322
3	0.896474608	0.852339858	0.704163605	0.803411240	0.753671075
4	0.637486478	0.950591793	0.608896833	0.739392272	0.901128426
5	0.875894931	0.496796610	0.757066594	0.826395494	0.476323772

Distance Between Cluster Centroids

Nearest Cluster	1	2	3	4	5
1	.	2.919082344	2.205364165	2.924274103	3.452409176
2	2.919082344	.	1.887712815	2.174543206	2.989259461
3	2.205364165	1.887712815	.	2.413029181	2.685589992
4	2.924274103	2.174543206	2.413029181	.	2.791329073
5	3.452409176	2.989259461	2.685589992	2.791329073	.

Table 6. Cluster means, cluster standard deviations and distance between cluster centroids of k-means clustering on HBAT customers using variables X6, X8, X12, X15 and X18 with 5 clusters.

Each cluster solution provides its strengths and weaknesses. More differentiation between groups is achieved using the four- and five-cluster approach while the three-cluster solution is more parsimonious. However, the five-cluster solution contains two clusters with low frequencies. Since clusters with fewer than 10% of observations are less favorable, I wouldn´t choose the five-cluster solution.
Investigating the cluster means of the four-cluster solution (Table 4, top) shows that variable X6 exhibits a relatively low score for cluster 3. Cluster 3 with 12 observations only, is further characterized by low cluster mean scores on variables X8 and X12. Cluster 4 has 25 observations but low scores for both X15 and X18. These results suggest that the four-cluster solution might also not be ideal for the given research question.
When looking at the results for the three-cluster solution (Table 2, top), one can make out more distinctive characteristics. The biggest cluster is cluster 3 with 64 observations. Variable X12 has the lowest cluster mean score for cluster 3. The low value for variable X8 characterizes cluster 2 which contains 12 observations. Cluster 1 with 24 observations is characterized by a relatively low score on X6.
Since the number of observations in the data set is only 100, I would choose the cluster solution with three clusters. Because increasing the number of clusters on a small data set doesn´t seem to yield representative cluster solutions.