DJILLALI LIABES UNIVERSITY OF SIDI BEL ABBES
FACULTY OF EXACT SCIENCES
DEPARTMENT OF COMPUTER SCIENCES

Module : Apprentissage Automatique
1ST YEAR OF MASTER'S DEGEREE IN
NETWORKS,SYSTEMS & INFORMATION SECURITY(RSSI)
2021/2022

Algorithmes Régression Linéaire simple et Kmeans

Student:
HADJAZI Mohammed Hisham
Group: 01 / RSSI

Module Instructor:
Pr.ELBERRICHI Zakaria
TD Instructor:
Dr.FAHSI.Mahmoud

A paper submitted in fulfilment of the requirements for the
Apprentissage Automatique TD-07

January 10, 2022

Contents

List of Figures

Chapter 1

Fiche TD-07 Solutions

1.1 Questions de cours.

1.1.1 1. Principe de la régression en général et de la régression linéaire en particulier ?

Regression is the statistical approach to find the relationship between variables. Hence, the Linear Regression assumes a linear relationship between variables. Depending on the number of input variables, the regression problem classified into :

1. Simple linear regression.
2. Multiple linear regression.

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Linear Regression

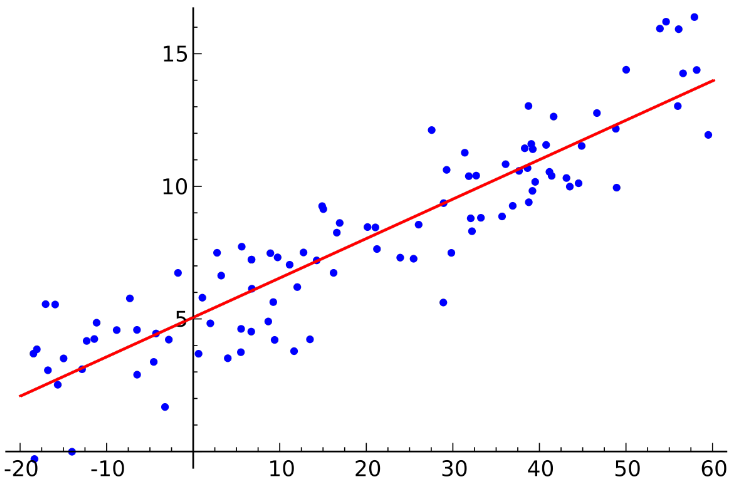

FIGURE 1.1: Linear Regression plot

Linear Regression is one of the most fundamental algorithms in Machine Learning you will ever encounter. Linear Regression involves fitting a linear function through the data which can be used to predict a continuous linear value like prices, stocks, houses, etc. A continuous value is anything that can be any real number. Regression involves predicting continuous real values and classification involves predicting classes.

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).

2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

The formula for a simple linear regression is:

$$\widehat{y} = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- β_0 is the **intercept**, the predicted value of **y** when the **x** is 0.
- β_1 is the regression coefficient – how much we expect **y** to change as **x** increases.
- **X** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

1.1.2 2. Principe du clustering en général, ses types, puis celui du K-means en particulier ?

Clustering is a machine learning technique that involves grouping similar data points together into so called clusters. Clustering is an unsupervised learning method commonly used in data science and other fields.

A more formal definition:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

KMeans is probably the most well-known of all the clustering algorithm. Its goal is to separate the data into K distinct non-overlapping subgroups (clusters) of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares.

K-Means clustering algorithm is defined as an unsupervised learning method having an iterative process in which the dataset are grouped into k number of predefined non-overlapping clusters or subgroups, making the inner points of the cluster as similar as possible while trying to keep the clusters at distinct space it allocates the data points to a cluster so that the sum of the squared distance between the clusters centroid and the data point is at a minimum, at this position the centroid of the cluster is the arithmetic mean of the data points that are in the clusters.

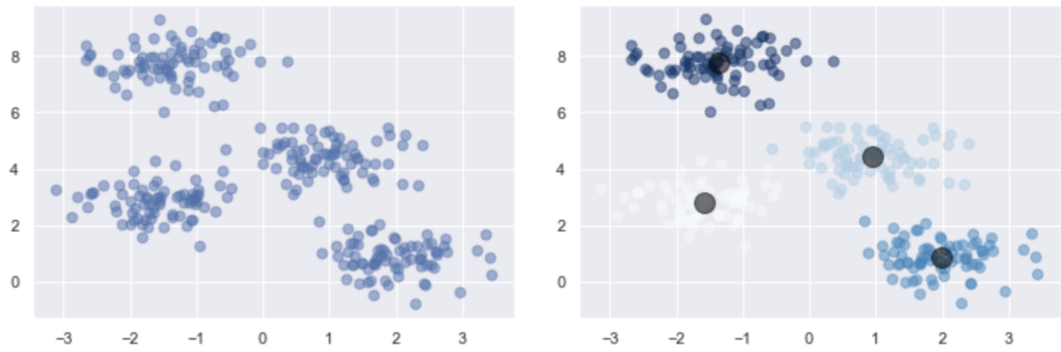

FIGURE 1.2: K-Means Clustering

This algorithm is an iterative algorithm that partitions the dataset according to their features into K number of predefined non- overlapping distinct clusters or subgroups. It makes the data points of inter clusters as similar as possible and also tries to keep the clusters as far as possible. It allocates the data points to a cluster if the sum of the squared distance between the cluster's centroid and the data points is at a minimum, where the cluster's centroid is the arithmetic mean of the data points that are in the cluster. A less variation in the cluster results in similar or homogeneous data points within the cluster.

K- Means Clustering Algorithm needs the following inputs:

- K = number of subgroups or clusters.
- Sample or Training Set = $\{x_1, x_2, x_3, ..., x_n\}$.

KMeans makes use of the sum-of-squares criterion, which works well if the clusters have a spherical-like shape. It doesn't work well on many other types of data like complicated shapes, though. In this section, we'll go over a few cases where KMeans performs poorly.

Also, as mentioned at the start of the section KMeans performs poorly for complicated geometric shapes such as the moons and circles shown below.

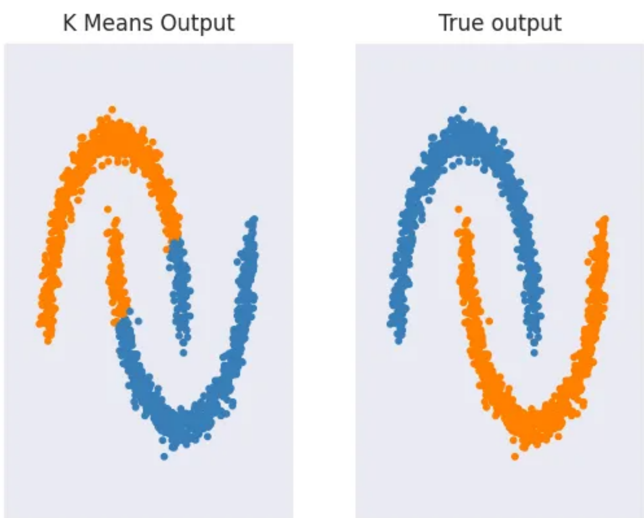

FIGURE 1.3: K-Means result 1

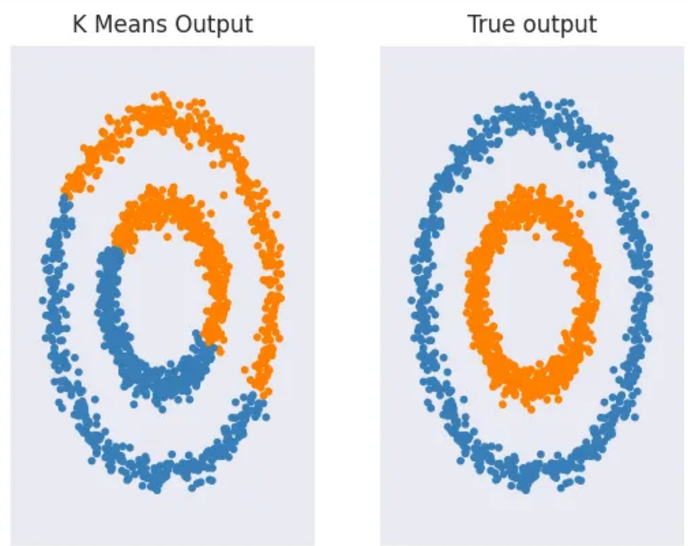

FIGURE 1.4: K-Means result 2

Other clustering algorithms like Spectral Clustering, Agglomerative Clustering, or DBSCAN don't have any problems with such data. For a more in-depth analysis of how different clustering algorithms perform on different interesting 2d dataset

1.1.3 3. Donner des exemples d'application des 2 algorithmes ?

Simple Linear Regression

1. **Marks scored by students based on number of hours studied** - Here marks scored in exams are independent and the number of hours studied is independent.

2. **Predicting crop yields based on the amount of rainfall** - Yield is a dependent variable while the measure of precipitation is an independent variable.

3. **Predicting the Salary of a person based on years of experience** - Therefore, Experience becomes the independent while Salary turns into the dependent variable.

4. **Market research studies and customer survey results analysis.**

5. **Studying engine performance from test data in automobiles.**

K- Means

1. **Academic Performance** - Based on the scores, students are categorized into grades like A, B, or C.

2. **Diagnostic systems** - The medical profession uses k-means in creating smarter medical decision support systems, especially in the treatment of liver ailments.

3. **Search engines** - Clustering forms a backbone of search engines. When a search is performed, the search results need to be grouped, and the search engines very often use clustering to do this.

4. **Wireless sensor networks** - The clustering algorithm plays the role of finding the cluster heads, which collect all the data in its respective cluster.
5. **Delivery Store Optimization** - Optimize the process of good delivery using truck drones by using a combination of k-means to find the optimal number of launch locations and a genetic algorithm to solve the truck route as a traveling salesman problem.
6. **Identifying Crime Localities** - With data related to crimes available in specific localities in a city, the category of crime, the area of the crime, and the association between the two can give quality insight into crime-prone areas within a city or a locality.
7. **Insurance Fraud Detection** - Machine learning has a critical role to play in fraud detection and has numerous applications in automobile, healthcare, and insurance fraud detection. utilizing past historical data on fraudulent claims, it is possible to isolate new claims based on its proximity to clusters that indicate fraudulent patterns. since insurance fraud can potentially have a multi-million dollar impact on a company, the ability to detect frauds is crucial.

1.2 Exercice 1 : En utilisant la régression linéaire simple (le critère des moindres carrés) sur la table de données suivante, donner l'équation Y = AX + B. Quelle valeur est prédite pour X = 180 ?

Valeurs de X (attribut)	155	170	160	185	195	190	175
Valeurs correspondantes observées de Y (classe)	45	70	65	80	85	95	60

								avarage	
Valeurs de X (attribut)	155	170	160	185	195	190	175	175.71	
Valeurs correspondantes observées de Y (classe)	45	70	65	80	85	95	60	71.42	
$(X - \overline{X})$	-20.71	-5.71	-15.71	9.28	19.28	14.28	-0.71		
$(Y - \overline{Y})$	-26.42	-1.42	-6.42	8.57	13.57	23.57	-11.42		
$(X - \overline{X})^2$	429.08	32.65	246.93	86.22	371.93	204.08	0.51	1371.42	b1
$(X - \overline{X})(Y - \overline{Y})$	547.44	8.16	101.02	79.59	261.73	336.73	8.16	1342.85	0.97

$$
\begin{gathered}
\widehat{y} = \beta_0 + \beta_1 X \\
71.4 = \beta_0 + 0.98(175.7) \\
\beta_0 = -100,786 \\
\widehat{y} = -100,786 + 0.98 * X \\
\widehat{y} = -100,786 + 0.98 * 180 \\
\widehat{y} = 75.614
\end{gathered}
$$

1.3 Exercice 2 : En utilisant l'algorithme K-Means, générer 2 clusters en prenant E7 et E8 comme centres initiaux et la distance de Manhattan, pour les données suivantes :

N exempls	ATT 1	ATT 2
E1	6	16
E2	5	17
E3	7	11
E4	5	14
E5	9	18
E6	6	17
E7	11	14
E8	4	10

1.3.1 First Iteration with Centroid E7 , E8

N exempls	ATT 1	ATT 2	CLUSTER 1	CLUSTER 2
E1	6	16	7	8
E2	5	17	9	8
E3	7	11	7	4
E4	5	14	6	5
E5	9	18	6	13
E6	6	17	8	9
E7	11	14	0	11
E8	4	10	11	0

CLUSTER 1	ATT 1	ATT 2	CLUSTER 2	ATT 1	ATT 2
E1	6	16	E2	5	17
E5	9	18	E3	7	11
E6	6	17	E4	5	14
E7	11	14	E8	4	10
Centroid 1	8	16.25	Centroid 2	5.25	13.00

1.3.2 Second Iteration with new Centroids (8,16.25) , (5.25,13)

N exempls	ATT 1	ATT 2	CLUSTER 1	CLUSTER 2
E1	6	16	2.25	3.75
E2	5	17	3.75	4.25
E3	7	11	6.25	3.75
E4	5	14	5.25	1.25
E5	9	18	2.75	8.75
E6	6	17	2.75	4.75
E7	11	14	5.25	6.75
E8	4	10	10.25	4.25

CLUSTER 1	ATT 1	ATT 2	CLUSTER 2	ATT 1	ATT 2
E1	6	16	E3	7	11
E2	5	17	E4	5	14
E5	9	18	E8	4	10
E6	6	17			
E7	11	14			
Centroid 1	7.4	16.4	Centroid 2	5.33	11.67

1.3.3 Third Iteration with new Centroids (7.4,16.4) , (5.33,11.67)

N exempls	ATT 1	ATT 2	CLUSTER 1	CLUSTER 2
E1	6	16	1.8	5
E2	5	17	3	5.66
E3	7	11	5.8	2.34
E4	5	14	4.8	2.66
E5	9	18	3.2	10
E6	6	17	2	6
E7	11	14	6	8
E8	4	10	9.8	3

CLUSTER 1	ATT 1	ATT 2	CLUSTER 2	ATT 1	ATT 2
E1	6	16	E3	7	11
E2	5	17	E4	5	14
E5	9	18	E8	4	10
E6	6	17			
E7	11	14			
Centroid 1	7.4	16.4	Centroid 2	5.33	11.67

Since no change after iteration 3 the final clusters are as follow :

1. **Cluster 1 :** E1,E2,E5,E6,E7
2. **Cluster 2 :** E3,E4,E8

1.4 Exercice 3 : En utilisant la régression linéaire simple (le critère des moindres carrés) sur la table de données suivante, donner la fonction Y = AX + B. Quelle valeur est prédite pour X = 35 ?

Valeur de X (attribut)	11.00	14.00	24.00	21.00	38.00	55.00	42.00
Valeurs correspondantes observées de Y (classe)	25.00	40.00	55.00	50.00	79.00	116.00	86.00

								avarage	
Valeur de X (attribut)	11.00	14.00	24.00	21.00	38.00	55.00	42.00	29.29	
Valeurs correspondantes observées de Y (classe)	25.00	40.00	55.00	50.00	79.00	116.00	86.00	64.43	
$(X-\overline{X})$	-18.29	-15.29	-5.29	-8.29	8.71	25.71	12.71		
$(Y-\overline{Y})$	-39.43	-24.43	-9.43	-14.43	14.57	51.57	21.57		
$(X-\overline{X})^2$	334.37	233.65	27.94	68.65	75.94	661.22	161.65	1563.43	B1
$(X-\overline{X})(Y-\overline{Y})$	720.98	373.41	49.84	119.55	126.98	1326.12	274.27	2991.14	1.91

$$\widehat{y} = \beta_0 + \beta_1 X$$
$$64.43 = \beta_0 + 1.91(29.29)$$
$$\beta_0 = 8.4861$$
$$\widehat{y} = 8.4861 + 1.91 * X$$
$$\widehat{y} = 8.4861 + 1.91 * 35$$
$$\widehat{y} = 75.3361$$

1.5 Exercice 4 : En utilisant la régression linéaire simple (le critère des moindres carrés) sur la table de données suivante, donner l'équation Y = AX + B. Quelle valeur est prédite pour X = 85 ?

Valeurs de X (attribut)	62	72	68	81	92	94	78
Valeurs correspondantes observées de Y (classe)	11.7	7.8	10.9	9	10.8	13.2	11.6

								Avarage	
Valeurs de X (attribut)	62	72	68	81	92	94	78	78.14	
Valeurs correspondantes observées de Y (classe)	11.7	7.8	10.9	9	10.8	13.2	11.6	11.16	
$(X - \overline{X})$	-16.14	-6.14	-10.14	2.86	13.86	15.86	-0.14		
$(Y - \overline{Y})$	0.54	-3.36	-0.26	-2.16	-0.36	2.04	0.44		
$(X - \overline{X})^2$	260.59	37.73	102.88	8.16	192.02	251.45	0.02	852.86	B1
$(X - \overline{X})(Y - \overline{Y})$	-8.66	20.66	2.67	-6.18	-5.03	32.30	-0.06	35.69	0.04184

$$\begin{gathered}
\widehat{y} = \beta_0 + \beta_1 X \\
11.16 = \beta_0 + 1.91(78.14) \\
\beta_0 = 7.89 \\
\widehat{y} = 7.89 + 0.04184 * X \\
\widehat{y} = 7.89 + 0.04184 * 85 \\
\widehat{y} = 11.4464
\end{gathered}$$

1.6 Utilisez l'algorithme k-means et la distance euclidienne pour regrouper les 8 exemples suivants en 3 clusters : A1(2,10) A2(2,5) A3(8,4) A4(5,8) A5(7,5) A6(6,4) A7(1,2) A8(4,9). On considère comme centre de classes à l'initialisation les points A1, A4 et A7.

1.6.1 First Iteration with Centroids A1, A4, A7

N exempls	ATT 1	ATT 2	CLUSTER 1	CLUSTER 2	CLUSTER 3
A1	2	10	0.00	3.61	8.06
A2	2	5	5.00	4.24	3.16
A3	8	4	8.49	5.00	7.28
A4	5	8	3.61	0.00	7.21
A5	7	5	7.07	3.61	6.71
A6	6	4	7.21	4.12	5.39
A7	1	2	8.06	7.21	0.00
A8	4	9	2.24	1.41	7.62

CLUSTER 1	ATT 1	ATT 2	CLUSTER 2	ATT 1	ATT 2	CLUSTER 3	ATT 1	ATT 2
A1	2	10	A3	8	4	A2	2	5
			A4	5	8	A7	1	2
			A5	7	5			
			A6	6	4			
			A8	4	9			
Centroid 1	2	10	Centroid 2	6	6	Centroid 3	1.5	3.5

1.6.2 Second Iteration with Centroids (3,9.5), (6.5,5.25), (1.5,3.5)

N exempls	ATT 1	ATT 2	CLUSTER 1	CLUSTER 2	CLUSTER 3
A1	2	10	0.00	5.66	6.52
A2	2	5	5.00	4.12	1.58
A3	8	4	8.49	2.83	6.52
A4	5	8	3.61	2.24	5.70
A5	7	5	7.07	1.41	5.70
A6	6	4	7.21	2.00	4.53
A7	1	2	8.06	6.40	1.58
A8	4	9	2.24	3.61	6.04

CLUSTER 1	ATT 1	ATT 2	CLUSTER 2	ATT 1	ATT 2	CLUSTER 3	ATT 1	ATT 2
A1	2	10	A3	8	4	A2	2	5
A8	4	9	A4	5	8	A7	1	2
			A5	7	5			
			A6	6	4			
Centroid 1	3	9.5	Centroid 2	6.5	5.25	Centroid 3	1.5	3.5

1.6.3 Third Iteration with Centroids (3.67,9), (7,4.33), (1.5,3.5)

N exempls	ATT 1	ATT 2	CLUSTER 1	CLUSTER 2	CLUSTER 3
A1	2	10	1.12	6.54	6.52
A2	2	5	4.61	4.51	1.58
A3	8	4	7.43	1.95	6.52
A4	5	8	2.50	3.13	5.70
A5	7	5	6.02	0.56	5.70
A6	6	4	6.26	1.35	4.53
A7	1	2	7.76	6.39	1.58
A8	4	9	1.12	4.51	6.04

CLUSTER 1	ATT 1	ATT 2	CLUSTER 2	ATT 1	ATT 2	CLUSTER 3	ATT 1	ATT 2
A1	2	10	A3	8	4	A2	2	5
A4	5	8	A5	7	5	A7	1	2
A8	4	9	A6	6	4			
Centroid 1	3.67	9.00	Centroid 2	7.00	4.33	Centroid 3	1.5	3.5

1.6.4 Forth Iteration with Centroids (3.67,9), (7,4.33), (1.5,3.5)

N exempls	ATT 1	ATT 2	CLUSTER 1	CLUSTER 2	CLUSTER 3
A1	2	10	1.95	7.56	6.52
A2	2	5	4.33	5.04	1.58
A3	8	4	6.61	1.05	6.52
A4	5	8	1.66	4.18	5.70
A5	7	5	5.20	0.67	5.70
A6	6	4	5.52	1.05	4.53
A7	1	2	7.49	6.44	1.58
A8	4	9	0.33	5.55	6.04

CLUSTER 1	ATT 1	ATT 2	CLUSTER 2	ATT 1	ATT 2	CLUSTER 3	ATT 1	ATT 2
A1	2	10	A3	8	4	A2	2	5
A4	5	8	A5	7	5	A7	1	2
A8	4	9	A6	6	4			
Centroid 1	3.67	9.00	Centroid 2	7.00	4.33	Centroid 3	1.5	3.5

Since no change after iteration 4 the final clusters are as follow :

1. **Cluster 1 :** A1,A4,A8
2. **Cluster 2 :** A3,A5,A6
3. **Cluster 3 :** A2,A7

1.7 Utilisez l'algorithme k-means et la distance manhatan pour regrouper les 8 exemples suivants en 3 clusters : A1(2,10,5) A2(2,5,3) A3(8,4,1) A4(5,8,7) A5(7,5,3) A6(6,4,7) A7(1,2,4) A8(4,9,5). On considère comme centre de classes à l'initialisation les points A1, A4 et A7.

1.7.1 First Iteration with Centroids A1, A4, A7

N exempls	ATT 1	ATT 2	ATT 3	CLUSTER 1	CLUSTER 2	CLUSTER 3
A1	2	10	5	0	7	10
A2	2	5	3	7	10	5
A3	8	4	1	16	13	12
A4	5	8	7	7	0	13
A5	7	5	3	12	9	10
A6	6	4	7	12	5	10
A7	1	2	4	10	13	0
A8	4	9	5	3	4	11

CLUSTER 1	ATT 1	ATT 2	ATT 3	CLUSTER 2	ATT 1	ATT 2	ATT 3	CLUSTER 3	ATT 1	ATT 2	ATT 3
A1	2	10	5	A4	5	8	7	A2	2	5	3
A8	4	9	5	A5	7	5	3	A3	8	4	1
				A6	6	4	7	A7	1	2	4
Centroid 1	3.00	9.50	3.33	Centroid 2	6.00	5.67	5.67	Centroid 3	3.67	3.67	2.67

1.7.2 Forth Iteration with Centroids (3,9.5,3.33), (6,5.67,5.67), (3.67,3.67;2.67)

N exempls	ATT 1	ATT 2	ATT 3	CLUSTER 1	CLUSTER 2	CLUSTER 3
A1	2	10	5	3.17	9	10.33
A2	2	5	3	5.83	7.34	3.33
A3	8	4	1	12.83	8.34	6.33
A4	5	8	7	7.17	4.66	9.99
A5	7	5	3	8.83	4.34	4.99
A6	6	4	7	12.17	3	6.99
A7	1	2	4	10.17	10.34	5.67
A8	4	9	5	3.17	6	7.99

CLUSTER 1	ATT 1	ATT 2	ATT 3	CLUSTER 2	ATT 1	ATT 2	ATT 3	CLUSTER 3	ATT 1	ATT 2	ATT 3
A1	2	10	5	A4	5	8	7	A2	2	5	3
A8	4	9	5	A5	7	5	3	A3	8	4	1
				A6	6	4	7	A7	1	2	4
Centroid 1	3.00	9.50	3.33	Centroid 2	6.00	5.67	5.67	Centroid 3	3.67	3.67	2.67

Since no change after iteration 2 the final clusters are as follow :

1. **Cluster 1 :** A1,A8
2. **Cluster 2 :** A4,A5,A6
3. **Cluster 3 :** A2,A3,A7

Bibliography

[1] Rebecca Bevans. *Simple Linear Regression, An Easy Introduction and Examples*. Scribbr, Feb. 2020. URL: https://www.scribbr.com/statistics/simple-linear-regression/.

[2] DZone. *10 Interesting Use Cases for the K-Means Algorithm - DZone AI*. dzone.com, May 2019. URL: https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-algorithm.

[3] Abhay Jidge. *The Complete Guide to Linear Regression Analysis*. Medium, Sept. 2020. URL: https://towardsdatascience.com/the-complete-guide-to-linear-regression-analysis-38a421a89dc2.

[4] *K- Means Clustering Algorithm, How It Works, Analysis and Implementation*. EDUCBA, May 2019. URL: https://www.educba.com/k-means-clustering-algorithm/.

[5] *K-Means Clustering Algorithm*. Simplilearn.com, Dec. 2021. URL: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm.

[6] Riya Kumari. *Simple Linear Regression: Applications, Limitations and Examples, Analytics Steps*. www.analyticssteps.com. URL: https://www.analyticssteps.com/blogs/simple-linear-regression-applications-limitations-examples.

[7] Gilbert Tanner. *KMeans Explained*. Machine Learning Explained, 2020. URL: https://ml-explained.com/blog/kmeans-explained.

[8] Vatsal. *K-Means Explained*. Medium, Feb. 2021. URL: https://towardsdatascience.com/k-means-explained-10349949bd10 (visited on 01/08/2022).

[9] Soner Yıldırım. *K-Means Clustering — Explained*. Medium, Mar. 2020. URL: https://towardsdatascience.com/k-means-clustering-explained-4528df86a120.