CLUSTERING

Dr. Debanjan Pathak

Introduction to Unsupervised Learning, Distance Metrics used

Unsupervised Learning

- It is a form of machine learning tasks where output $y^{(i)}$ is not available in training data.
- In this type of problem, we are given a set of feature vectors $x^{(i)}$, and the goal is to unravel the underlying similarities.
- Unsupervised learning is more about creative endevours—exploration, understanding and refinements that do not lend themselves to a specific set of steps, i.e, there is no specific methodology.
- There is no right or wrong answer; no simple statistical measure that summarizes the goodness of results.
- Instead, descriptive statistics and visualization are key parts of the process.

Unsupervised Learning

- The requirement of dataset partitioning—into training, validation and test sets—is of little importance.
- There are two types:
  - **Cluster Analysis**
  - **Association Rules**
- Use the labeled data to train a classifier.
- The next step is to apply the same to the unlabeled data so that it is labeled with class probabilities.
- The third step is to train another classifier with the help of labels for all the data.
- Fourth, repeat till convergence (the termination of the optimization algorithm) is achieved.

Cluster Analysis

- Cluster analysis is employed to create groups or clusters **of similar records** on the basis of many measurements made for these records.
- A primary issue in clustering is that of defining ‘similarity’ between feature vectors $x^{(i)}; i = 1, 2, \ldots, N,$ representing the records.
- Another important issue is the selection of an algorithmic scheme that will cluster (group) the vectors based on the accepted similarity measure.
- The level of similarity and dissimilarity are evaluated on the basis of the characteristics of the variables that describe the objects or components.
- This assessment often involves distance measures.

Euclidean distance
Statistical distance
Manhattan distance
Minkowski metric
Hamming distance

Already discussed in previous class

Cluster Analysis

- Clustering can be used for data exploration, to understand the structure of the data.
- A natural way to make sense of complex data is to break the data into smaller clusters of data; then finding patterns within each cluster is often possible.
- Another use of clustering methods is in outlier detection which is finding instances that do not lie in any of the main clusters and are exceptions.
- The outliers may be recording errors that should be detected and discarded in data cleansing process.
- While selecting the right clustering algorithm, we should make use of the knowledge of the problem the dataset describes.

Cluster Analysis

- Data partition must usually have two features:
  - Clusters should be homogeneous within: **Data within each cluster should strongly resemble each other.**
- Compactness is indicated by the variance of patterns in a cluster.
  - There should be heterogeneity between clusters: **Data should differ from one cluster to another cluster as much as possible.**
- The Euclidean distance between cluster centroids indicates the cluster separation.

Association Rules

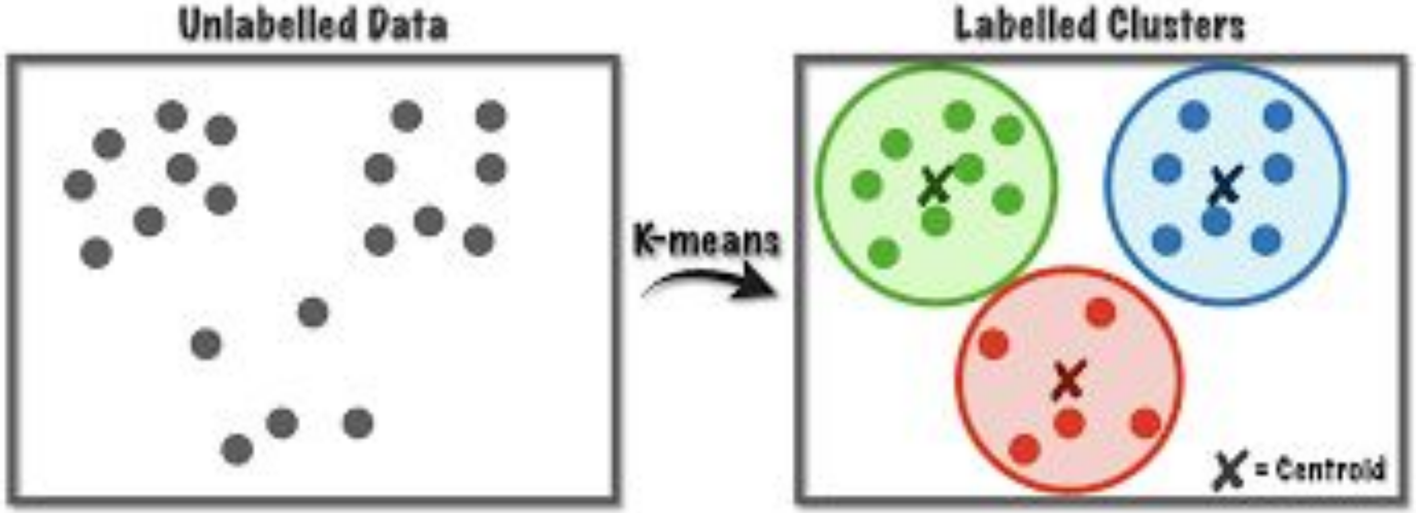
- Frequent patterns detection and generating association rules is another application of unsupervised learning.
- Mining frequent patterns is originated with the study of customer transactions databases to determine association between purchases of different items/service offerings.
- This popular area of application is called market basket analysis, which studies customers’ buying habits for products that are purchased together.
- This application is commonly encountered in online recommender systems where customers examining product(s) for possible purchase are shown other products that are frequently purchased in conjunction with the desired product(s); display from Amazon.com, for example.

Association Rules

- Other than market basket data, association analysis can also be applied to web mining, medical diagonosis, text mining, scientific data analysis, and other application domains.
- A medical researcher wishes to find out about the symptoms that match the confirmed diagonosis.
- While analysing Earth science data, the association patterns often disclose interesting links among the ocean, land and atmospheric pressures.
- The association approach, in terms of text documents, can be used to discover word co-occurrence relationships (used to find linguistic patterns).
- Association analysis can be used to discover web usage patterns.

K-Means Approach for Clustering

Basics



- The most general of the heuristic clustering techniques is the K-means clustering.
- It is amongst the widely used clustering algorithms.
- K-means clustering characteristically espouses exclusive cluster separation:
    - The set of **all clusters comprises all data vectors**.
    - Each object **belongs to exactly one group**.
    - **None of the clusters is empty and none of them contain the entire dataset X**. The clusters are not joined.

Image compression using **clustering**

- In an image, if we decide to color code shades of the same group with a single color, say their average, then we are actually quantizing the image.
- If 24 bit pixels represent 16 million colors for an image, and there are shades of merely 64 main colors, we will require 6 bits for each pixel rather than 24.

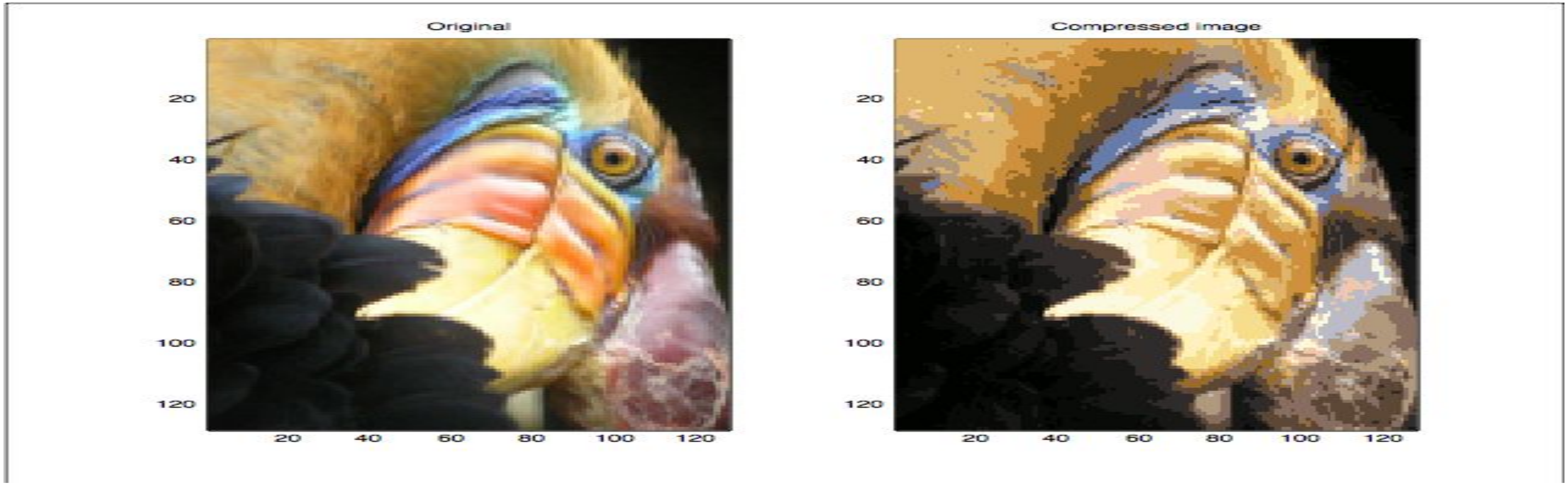


Image compression using **clustering**

- K-means can be applied on the pixel values to get the resultant compressed image.
- Now, these ‘k’ cluster centroids will replace all the color vectors in their respective clusters.
- we have only reduced the number of colors to represent a colored digital image known as Color Quantization.
- However, there can be several different ways to achieve this target.
- Few other methods can be reducing the size of the image or reducing the intensity ranges of pixels or reducing the frequency of an image.

How does it work

- The goal of clustering algorithm is to find K points that make good cluster centers.
- These centers define the clusters.
- Each object is assigned to the cluster defined by the nearest cluster center.
- The best assignment of cluster centers could be defined as the one that minimizes the sum of distances (or the distance-squared) from every point to its nearest cluster center.
- K points are randomly selected as cluster centers, which gives us cluster seeds.

Cluster Interpretability

- The characteristics of each cluster to be able to understand the clusters by:
  - first getting the summary statistics from each cluster;
  - then, establishing if the dataset provided has a nonrandom structure, which may result in meaningful clusters;
  - and finally attempting to allocate a name or label to each cluster.
- The most simple and commonly used criterion function for clustering is the sum-of-squared-error criterion.
- Let N_k be the no. of samples in cluster k and μ_k be the mean of those samples:

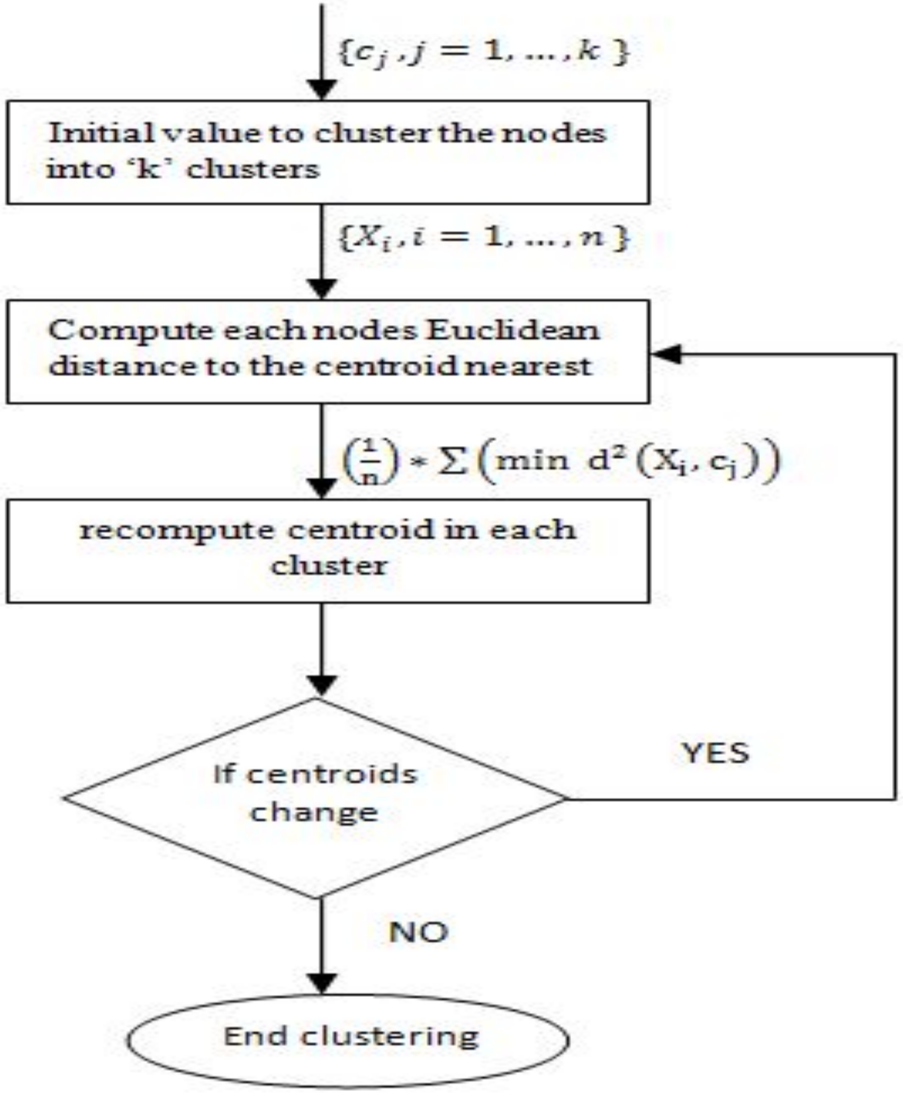
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{x}^{(i)}$$

Then the sum of squared errors is defined by,

$$J_e = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_k \right\|^2$$

where K stands for number of data partitionings.

Flow Chart



$\{c_j, j = 1, \dots, k\}$

Initial value to cluster the nodes into 'k' clusters

$\{X_i, i = 1, \dots, n\}$

Compute each nodes Euclidean distance to the centroid nearest

$\left(\frac{1}{n}\right) * \Sigma\left(\min\ d^2\left(X_i, c_j\right)\right)$

recompute centroid in each cluster

If centroids change

YES

NO

End clustering

Procedure

- All objects are allocated to their nearest cluster center as per the Euclidean distance metric.
- allocation of all objects to the nearest seed, all that is required is the calculation of the distance between each object and each seed.
- Then comes the calculation of the centroid or mean of the objects in each cluster, which is the ‘means’ part of the algorithm.
- These centroids are considered as new cluster-center values for their respective clusters.
- The entire procedure is iterated with the new cluster centers.
- Repetition goes on until the same points are assigned to each cluster in successive rounds.
- At this stage, the cluster centers become stable.

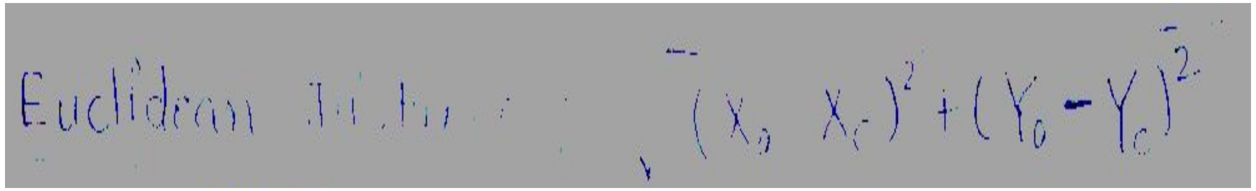
Points to be understood

- Entirely different arrangements can be derived from small changes in primary random selection of the seeds, and some may be more beneficial than others.
- To increase the chance of discovering the solution that has the maximum benefit, the algorithm **may require to be run many times with varioius seeds and then the best final result may be selected.**
- Similar situation arises in the choice of K.
- Often, nothing is known about the likely number of clusters, and the whole point of clustering is to find it out.
- A heuristic way is to try different values and choose the best.
- Try normalizing/standardizing the data as it involves with distance measures.

Performance Evaluation, and Stopping Criteria for K Means

Form two clusters for given data.

- We need to form 2 clusters, So for that we consider two data points of our data (randomly or based on i value) and assign them as a centroid for each cluster.
- Now we need to assign each and every data point of our data to one of these clusters based on Euclidean distance calculation.



Euclidean distance = $\sqrt{(X_0 - X_c)^2 + (Y_0 - Y_c)^2}$

- Here (X0,Y0) is our data point and (Xc,Yc) is a centroid of a particular cluster.

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Form two clusters for given data.

- Lets consider the 2nd data point i.e. (170,56) and check its distance with the centroid of both clusters.

$$K_1 = \sqrt{(170-185)^2+(56-72)^2}$$

$$= 21.93$$

$$K_2 = \sqrt{(170-180)^2+(56-71)^2}$$

$$= 18.03$$

- Now we can see from calculations that 2nd data point(170,56) is more closer to k2(cluster 2), so we assign it to k2.
- We shall continue this procedure for all data points.

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Form two clusters for given data

Height	Weight	K1 (185,72)	K2 (180,71)	Closer Centroid
185	72	0.00	5.10	K1
170	56	21.93	18.03	K2
168	60	20.81	16.28	K2
179	68	7.21	3.16	K2
182	72	3.00	2.24	K2
188	77	5.83	10.00	K1
180	71	5.10	0.00	K2
180	70	5.39	1.00	K2
183	84	12.17	13.34	K1
180	88	16.76	17.00	K1
180	67	7.07	4.00	K2
177	76	8.94	5.83	K2

1st iteration

$$K_{1\_1} \text{ (Height)} = \frac{185+188+183+180}{4} = 184$$

$$K_{1\_1} \text{ (weight)} = \frac{72+77+84+88}{4} = 80.25$$

$$K_{2\_1} \text{ (Height)} = \frac{180+170+168+\ldots+177}{8} = 177$$

$$K_{2\_1} \text{ (weight)} = \frac{71+56+60+\ldots+76}{8} = 67.25$$

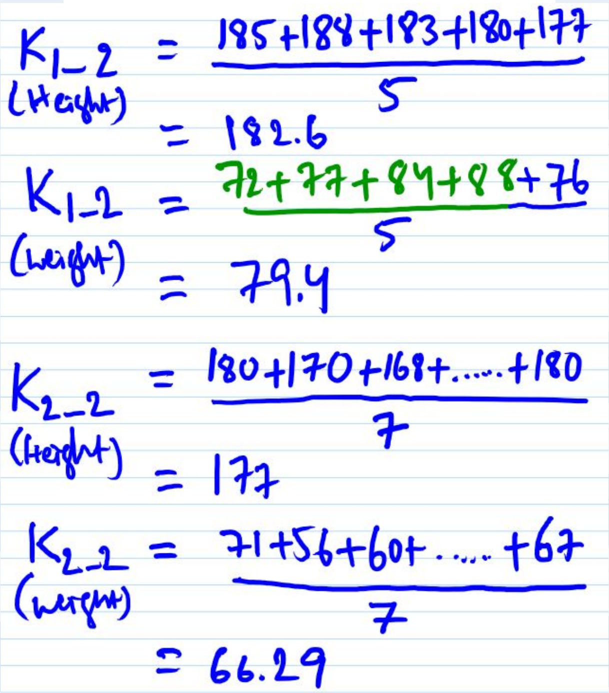
Height	Weight	k1_1 (184,80.25)	k2_1 (177,67.25)	Closer Centroid
185	72	8.31	9.18	K1
188	77	5.15	14.53	K1
183	84	3.88	17.56	K1
180	88	8.72	20.72	K1
180	71	10.08	4.61	K2
170	56	28.00	13.46	K2
168	60	25.81	11.72	K2
179	68	13.23	2.06	K2
182	72	8.49	6.73	K2
180	70	11.00	3.91	K2
180	67	13.84	3.04	K2
177	76	8.19	8.50	K1

https://docs.google.com/spreadsheets/d/1xOFlBp_sV-WnNcIz6VvjNaw-Hr4sw8M1/edit?usp=sharing &ouid=118186460474919932244&rtpof=true&sd=true

Form two clusters for given data

Height	Weight	k1_1 (184,80.25)	k2_1 (177,67.25)	Closer Centroid
185	72	8.31	9.18	K1
188	77	5.15	14.53	K1
183	84	3.88	17.56	K1
180	88	8.72	20.72	K1
180	71	10.08	4.61	K2
170	56	28.00	13.46	K2
168	60	25.81	11.72	K2
179	68	13.23	2.06	K2
182	72	8.49	6.73	K2
180	70	11.00	3.91	K2
180	67	13.84	3.04	K2
177	76	8.19	8.50	K1

2nd iteration



$$K_{1\_2}\ (\text{Height}) = \frac{185+188+183+180+177}{5} = 182.6$$

$$K_{1\_2}\ (\text{Weight}) = \frac{72+77+84+88+76}{5} = 79.4$$

$$K_{2\_2}\ (\text{Height}) = \frac{180+170+168+\ldots\ldots+180}{7} = 177$$

$$K_{2\_2}\ (\text{weight}) = \frac{71+56+60+\ldots\ldots+67}{7} = 66.29$$

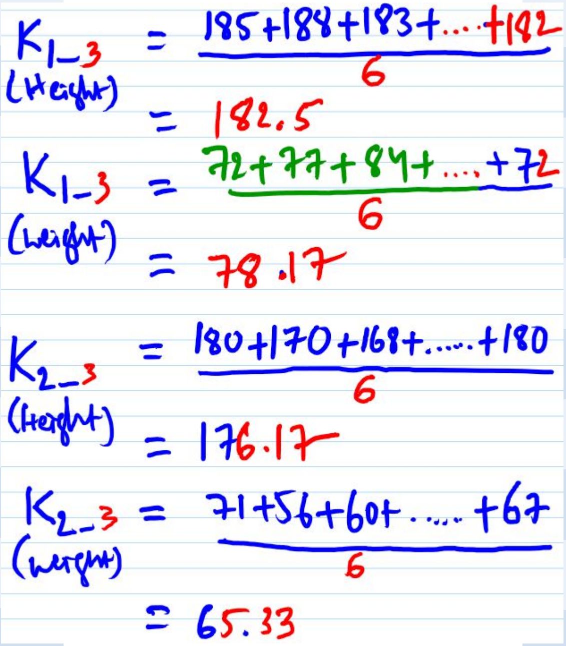
Height	Weight	k1_2 (182.6,79.4)	k2_2 (177,66.29)	Closer Centroid
185	72	7.78	9.83	K1
188	77	5.91	15.36	K1
183	84	4.62	18.70	K1
180	88	8.98	21.92	K1
177	76	6.55	9.71	K1
180	71	8.79	5.59	K2
170	56	26.58	12.44	K2
168	60	24.28	10.98	K2
179	68	11.95	2.63	K2
182	72	7.42	7.59	K1
180	70	9.75	4.77	K2
180	67	12.67	3.08	K2

https://docs.google.com/spreadsheets/d/1xOFlBp_sV-WnNcIz6VvjNaw-Hr4sw8M1/edit?usp=sharing
&ouid=118186460474919932244&rtpof=true&sd=true

Form two clusters for given data

Height	Weight	k1_2	k2_2	Closer
		(182.6,79.4)	(177,66.29)	Centroid
185	72	7.78	9.83	K1
188	77	5.91	15.36	K1
183	84	4.62	18.70	K1
180	88	8.98	21.92	K1
177	76	6.55	9.71	k1
180	71	8.79	5.59	K2
170	56	26.58	12.44	K2
168	60	24.28	10.98	K2
179	68	11.95	2.63	K2
182	72	7.42	7.59	K1
180	70	9.75	4.77	K2
180	67	12.67	3.08	K2

3rd iteration



$$K_{1\_3} \text{ (Height)} = \frac{185+188+183+\ldots.+182}{6}$$

$$= 182.5$$

$$K_{1\_3} \text{ (Weight)} = \frac{72+77+84+\ldots.+72}{6}$$

$$= 78.17$$

$$K_{2\_3} \text{ (Height)} = \frac{180+170+168+\ldots..+180}{6}$$

$$= 176.17$$

$$K_{2\_3} \text{ (weight)} = \frac{71+56+60+\ldots..+67}{6}$$

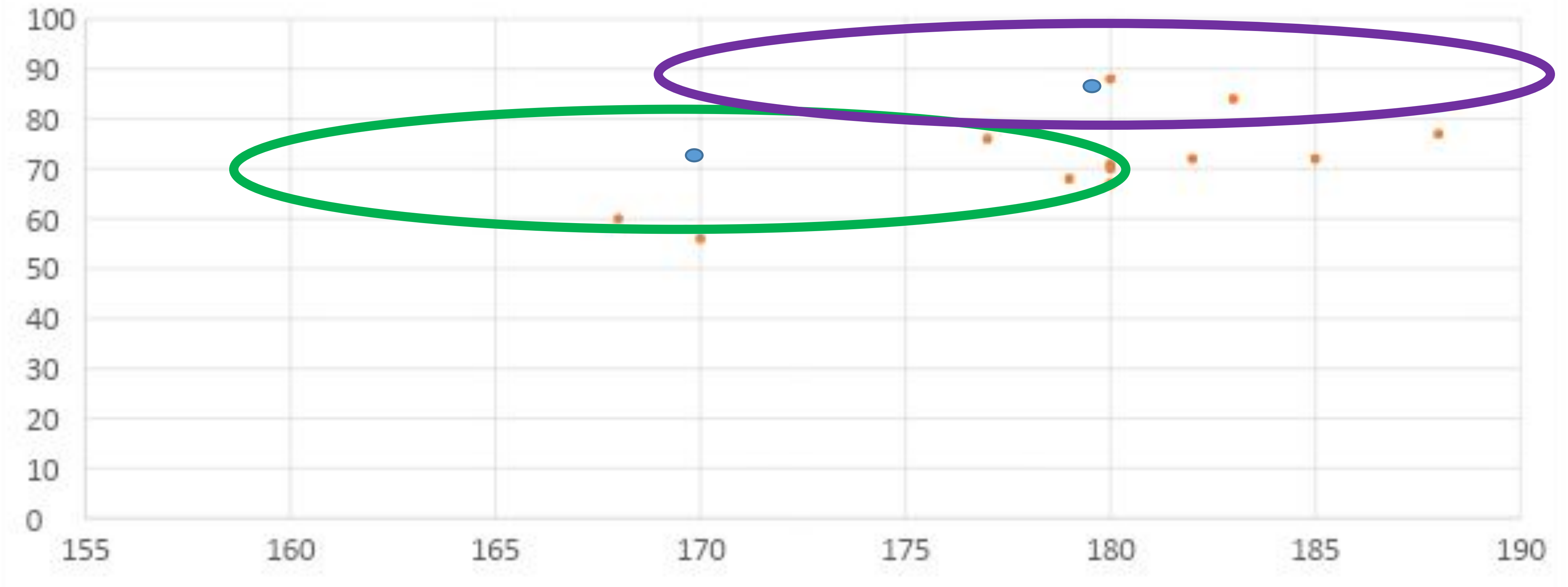
$$= 65.33$$

Height	Weight	k1_3	k2_3	Closer
		(182.5,78.17)	(176.17,65.33)	Centroid
185	72	6.65	11.07	K1
188	77	5.62	16.62	K1
183	84	5.85	19.88	K1
180	88	10.15	22.99	K1
177	76	5.91	10.70	K1
182	72	6.19	8.86	K1
180	71	7.59	6.84	K2
170	56	25.45	11.19	K2
168	60	23.24	9.75	K2
179	68	10.75	3.89	K2
180	70	8.54	6.04	K2
180	67	11.44	4.18	K2

https://docs.google.com/spreadsheets/d/1xOFlBp_sV-WnNcIz6VvjNaw-Hr4sw8M1/edit?usp=sharing
&ouid=118186460474919932244&rtpof=true&sd=true

How to know your result is correct or not?

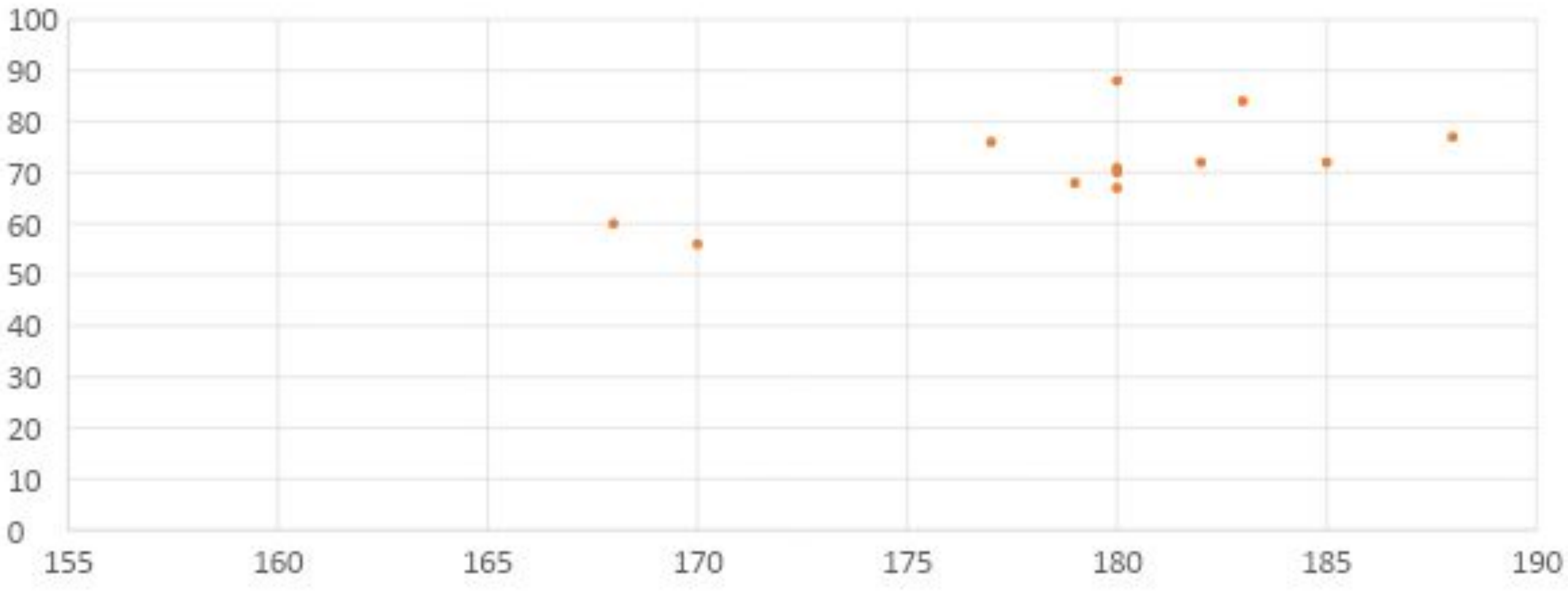
k1_3 (182.5,78.17) k2_3 (176.17,65.33)



Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

How to know your result is correct or not?

- Change the initial centroids; or chose them on the basis of their appearance in the graph.



Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Form two clusters for given data

Height	Weight	K1 (168,60)	K2 (180,71)	Closer Centroid
185	72	20.81	5.10	K2
170	56	4.47	18.03	K1
168	60	0.00	16.28	K1
179	68	13.60	3.16	K2
182	72	18.44	2.24	K2
188	77	26.25	10.00	K2
		178.39		
180	71	16.28	0.00	K2
180	70	15.62	1.00	K2
183	84	28.30	13.34	K2
180	88	30.46	17.00	K2
180	67	13.89	4.00	K2
177	76	18.36	5.83	K2

1st iteration

Height	Weight	k1_1 (169,58)	k2_1 (181,74.5)	Closer Centroid
170	56	2.24	21.73	K1
168	60	2.24	19.74	K1
185	72	21.26	4.38	K2
179	68	14.14	6.93	K2
182	72	19.10	2.57	K2
188	77	26.87	7.06	K2
180	71	17.03	3.77	K2
180	70	16.28	4.71	K2
183	84	29.53	9.63	K2
180	88	31.95	13.57	K2
180	67	14.21	7.63	K2
177	76	19.70	4.65	K2

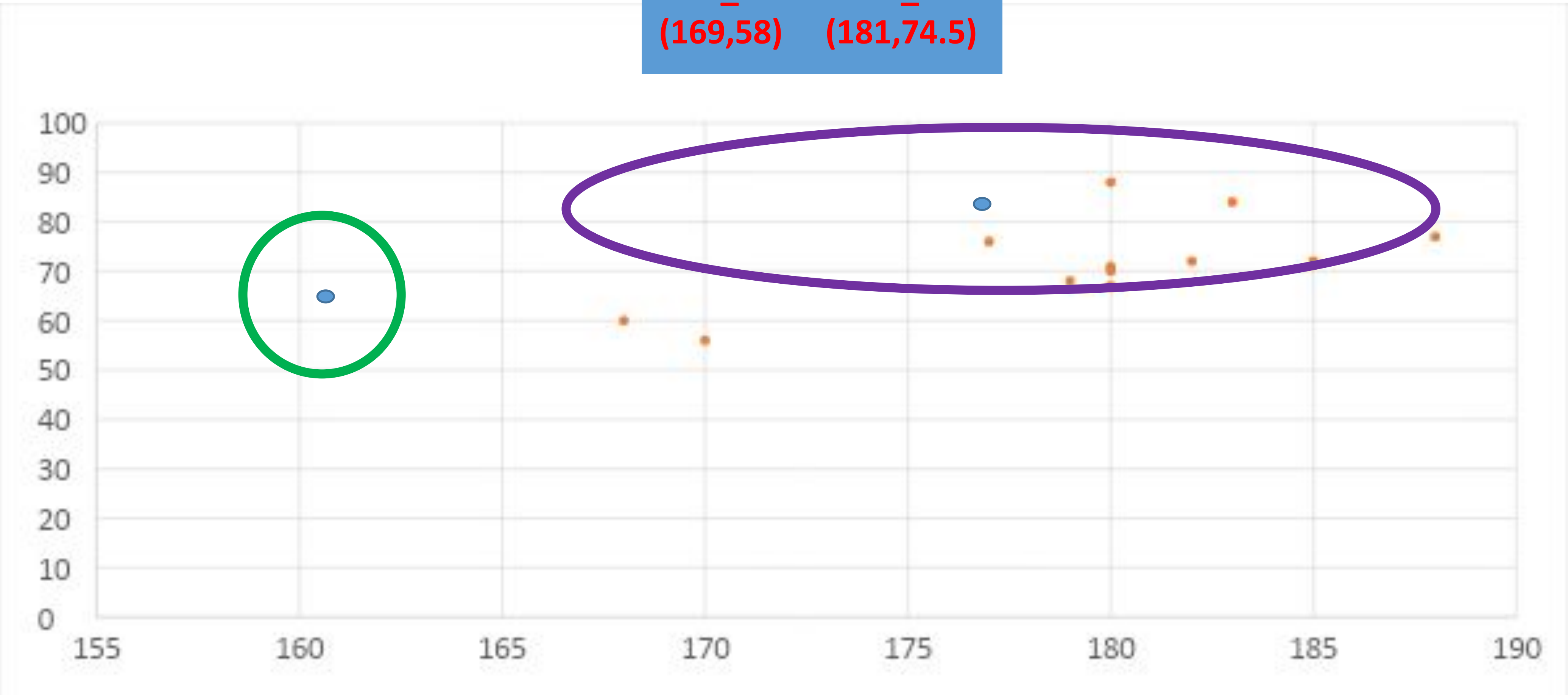
Stopping Criteria: No Change in centroids (happens when no exchange of data points takes place between clusters.

https://docs.google.com/spreadsheets/d/1xOFlBp_sV-WnNcIz6VvjNaw-Hr4sw8M1/edit?usp=sharing &ouid=118186460474919932244&rtpof=true&sd=true

How to know your result is correct or not?

- New Clusters:

k1_1	k2_1
(169,58)	(181,74.5)



Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Points to be understood (once again)

- Entirely different arrangements can be derived from small changes in primary random selection of the seeds, and some may be more beneficial than others.
- To increase the chance of discovering the solution that has the maximum benefit, the algorithm may require to be run many times with varioius seeds and then the best final result may be selected.
- Similar situation arises in the choice of K.
- Often, nothing is known about the likely number of clusters, and the whole point of clustering is to find it out.
- A heuristic way is to try different values and choose the best.
- Try normalizing/standardizing the data as it involves with distance measures.