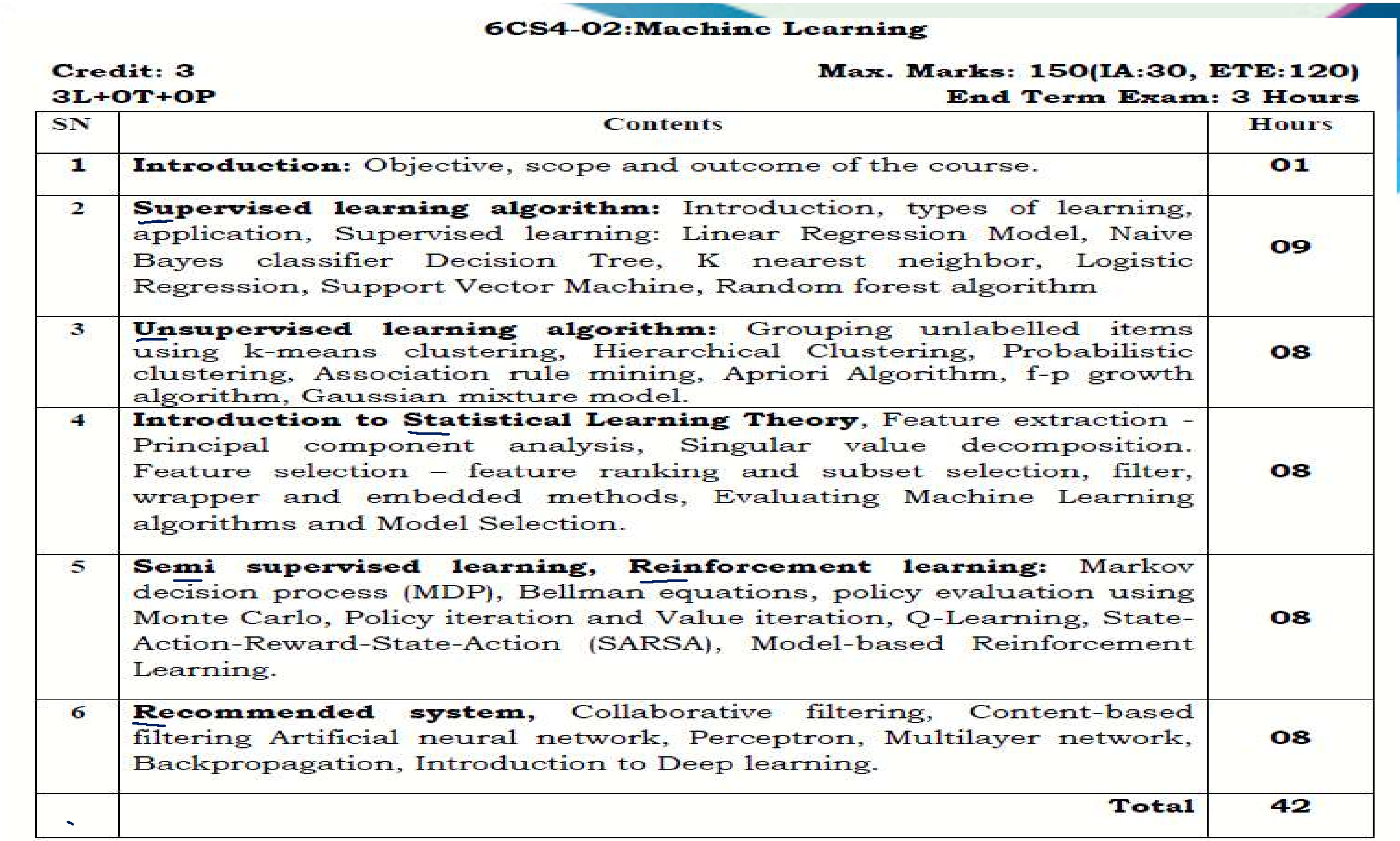


6CS4-02:Machine Learning

Credit: 3 **Max. Marks: 150(IA:30, ETE:120)**
3L+0T+0P **End Term Exam: 3 Hours**

SN	Contents	Hours
1	**Introduction:** Objective, scope and outcome of the course.	01
2	**Supervised learning algorithm:** Introduction, types of learning, application, Supervised learning: Linear Regression Model, Naive Bayes classifier Decision Tree, K nearest neighbor, Logistic Regression, Support Vector Machine, Random forest algorithm	09
3	**Unsupervised learning algorithm:** Grouping unlabelled items using k-means clustering, Hierarchical Clustering, Probabilistic clustering, Association rule mining, Apriori Algorithm, f-p growth algorithm, Gaussian mixture model.	08
4	**Introduction to Statistical Learning Theory**, Feature extraction - Principal component analysis, Singular value decomposition. Feature selection – feature ranking and subset selection, filter, wrapper and embedded methods, Evaluating Machine Learning algorithms and Model Selection.	08
5	**Semi supervised learning, Reinforcement learning:** Markov decision process (MDP), Bellman equations, policy evaluation using Monte Carlo, Policy iteration and Value iteration, Q-Learning, State-Action-Reward-State-Action (SARSA), Model-based Reinforcement Learning.	08
6	**Recommended system,** Collaborative filtering, Content-based filtering Artificial neural network, Perceptron, Multilayer network, Backpropagation, Introduction to Deep learning.	08
	Total	**42**

Machine Learning

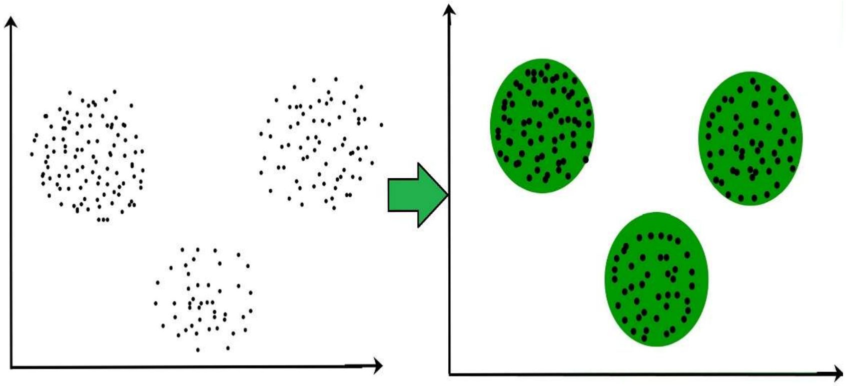
Unit-2

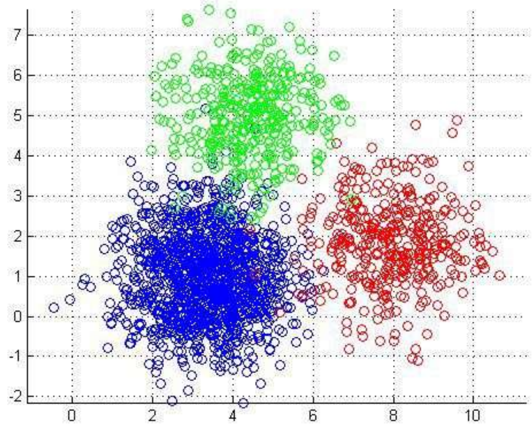
Unsupervised Learning Algorithm:

B K BIRLA INSTITUTE OF ENGINEERING & TECHNOLOGY, PILANI

Clustering:

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters





K-means clustering:

In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

Working of K-Means Algorithm

We can understand the working of K-Means clustering algorithm with the help of following steps −

Step 1 − First, we need to specify the number of clusters, K, need to be generated by this algorithm.

Step 2 − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.

Step 3 − Now it will compute the cluster centroids.

K-means clustering:

Step 4 − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more

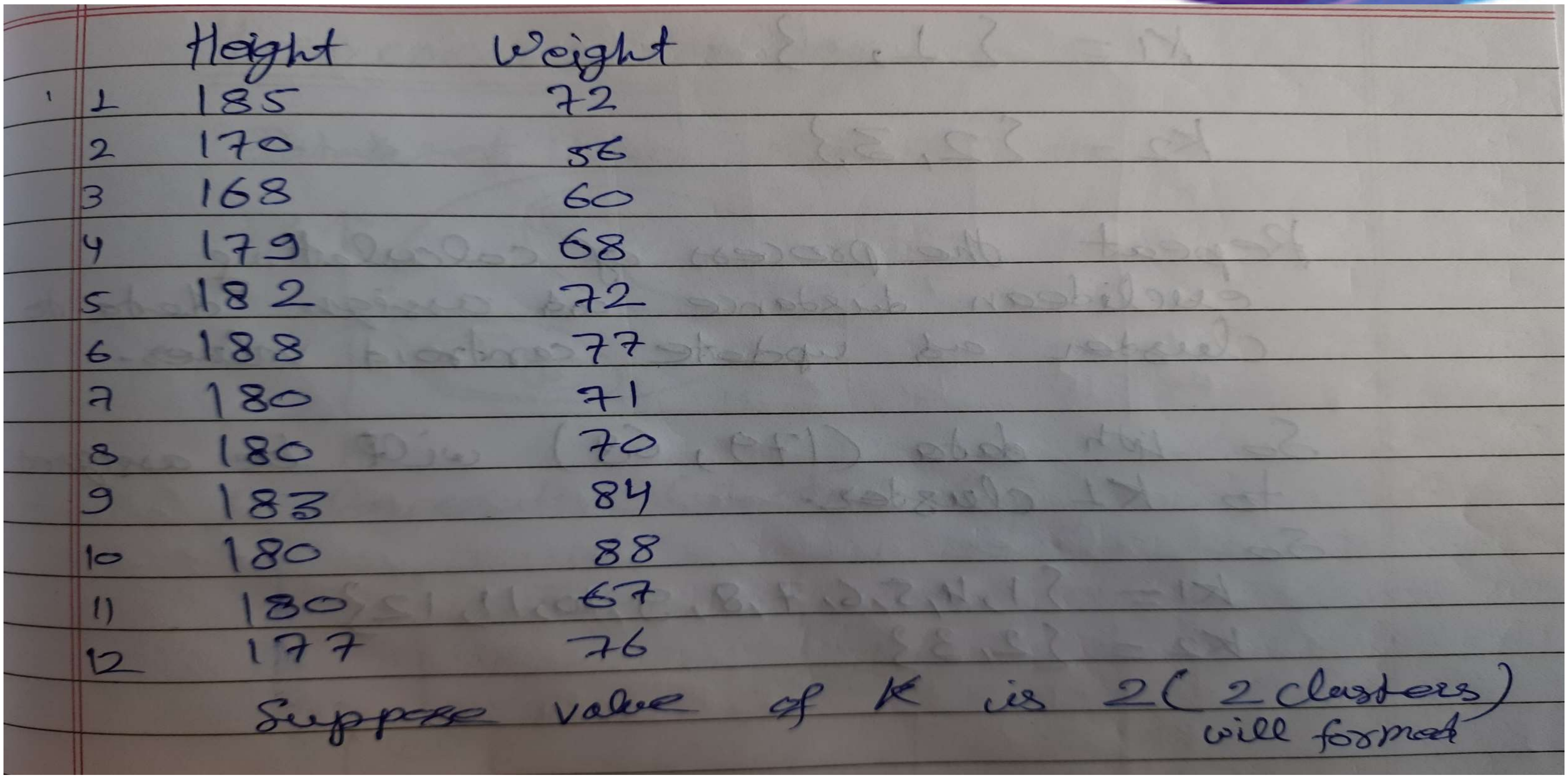
•**4.1** − First, the sum of squared distance between data points and centroids would be computed.

•**4.2** − Now, we have to assign each data point to the cluster that is closer than other cluster (centroid).

•**4.3** − At last compute the centroids for the clusters by taking the average of all data points of that cluster.

K-means follows **Expectation-Maximization** approach to solve the problem. The Expectation-step is used for assigning the data points to the closest cluster and the Maximization-step is used for computing the centroid of each cluster.

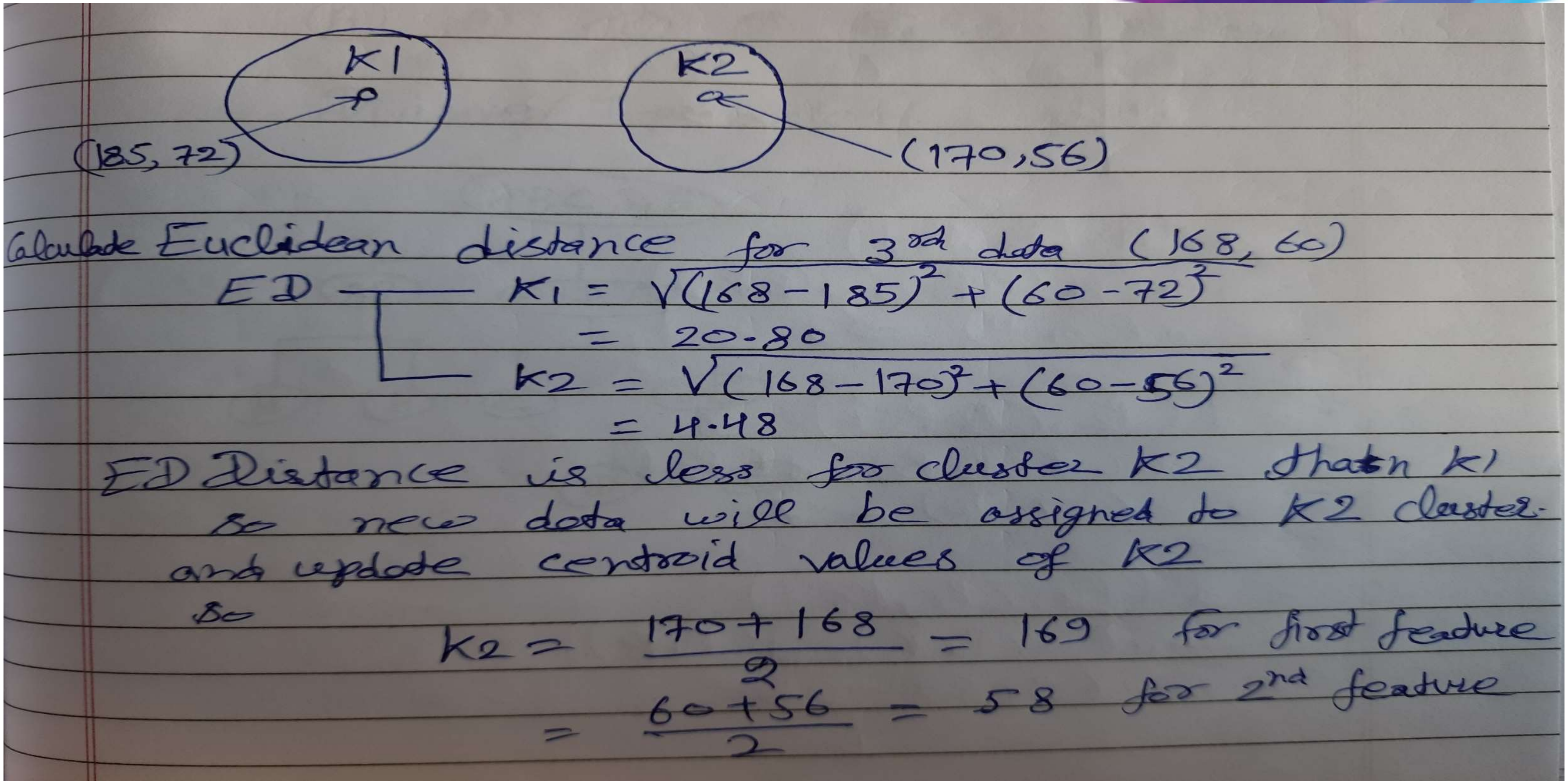
K-means clustering:



	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

Suppose value of K is 2 (2 clusters will formed)

K-means clustering:


K1
(185, 72)
K2
(170, 56)

Calculate Euclidean distance for 3rd data (168, 60)

ED

$$K_1 = \sqrt{(168-185)^2 + (60-72)^2}$$
$$= 20.80$$
$$K_2 = \sqrt{(168-170)^2 + (60-56)^2}$$
$$= 4.48$$

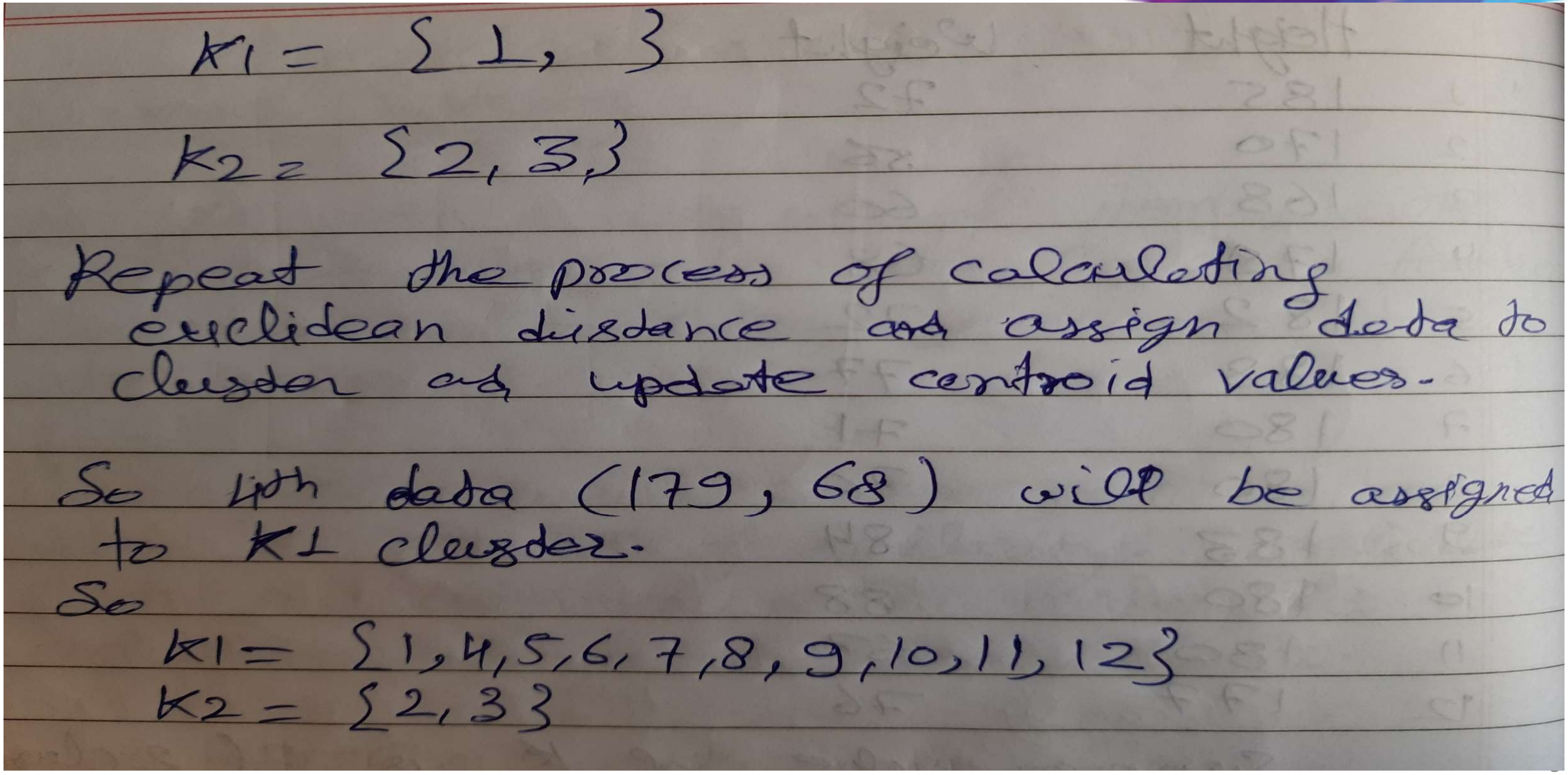
ED Distance is less for cluster K2 tha~~t~~n K1 so new data will be assigned to K2 cluster. and update centroid values of K2

so

$$K_2 = \frac{170+168}{2} = 169$$ for first feature

$$= \frac{60+56}{2} = 58$$ for 2nd feature

K-means clustering:



$K1 = \{1, \}$

$K2 = \{2, 3,\}$

Repeat the process of calculating euclidean distance and assign data to cluster and update centroid values.

So 14th data (179, 68) will be assigned to K1 cluster.

So

$K1 = \{1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

$K2 = \{2, 3\}$