upGrad

PLACEMENT REFRESHER PROGRAM

Session 9 - ML 3

Regression & Clustering

By
Ritesh Kumar Pandey

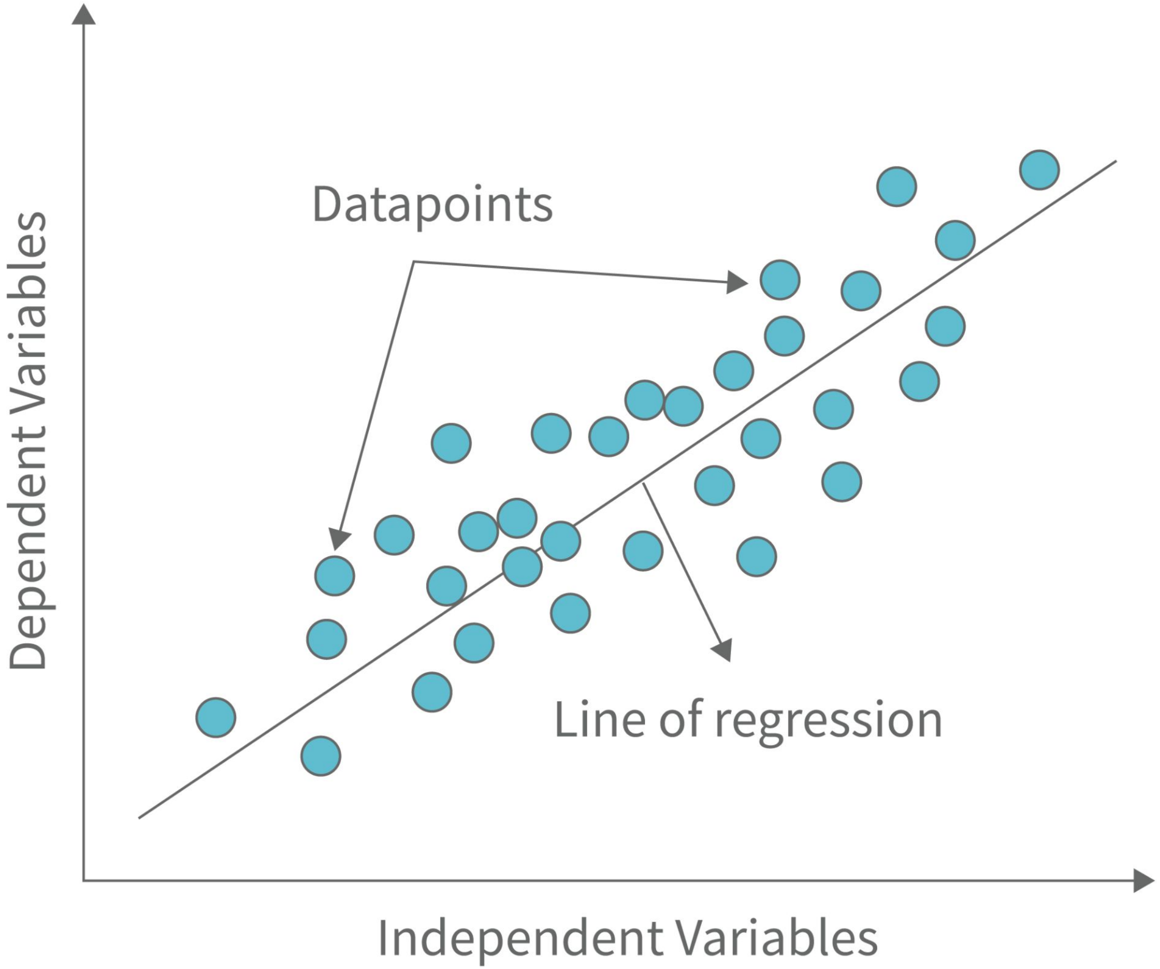
Agenda

- Linear Regression
- Cluster Analysis

Linear Regression

upGrad

- Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables.
- It is called "linear" because the model assumes a linear relationship between the dependent and independent variables.



Linear Regression - Equation

upGrad

Straight-line regression analysis involves a response variable, y, and a single predictor variable, x. It is the simplest form of regression, and models y as a linear function of x.

That is,

$$y = b + wx;$$

where the variance of y is assumed to be constant, and b and w are regression coefficients specifying the Y-intercept and slope of the line, respectively.

The regression coefficients, w and b, can also be thought of as weights, so that we can equivalently write,

$$y = w0 + w1x$$

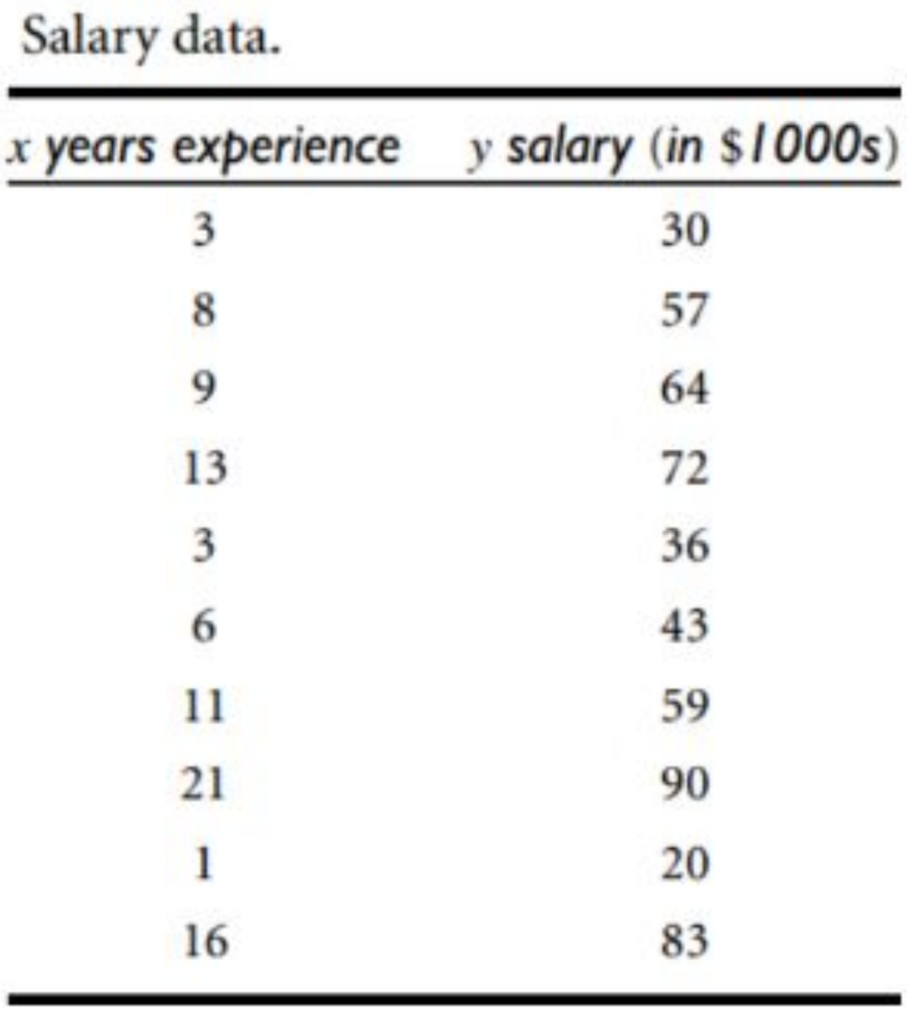
Linear Regression - Equation

The regression coefficients can be estimated using this method with the following equations:

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1\bar{x}$$

Linear Regression - Question

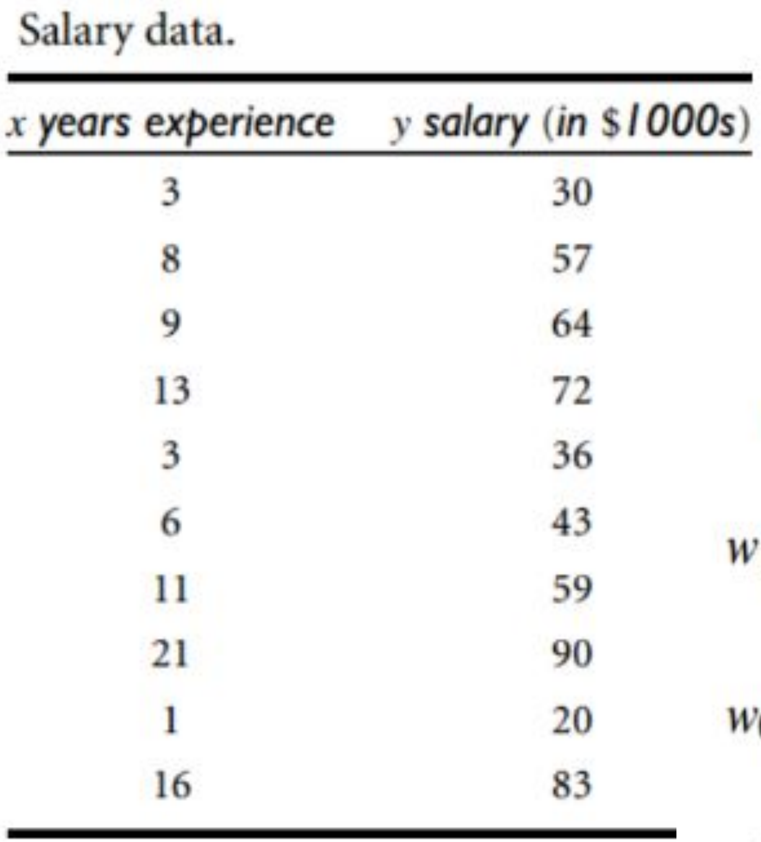


Salary data.

x years experience	y salary (in $1000s)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Form the linear regression equation for the given dataset

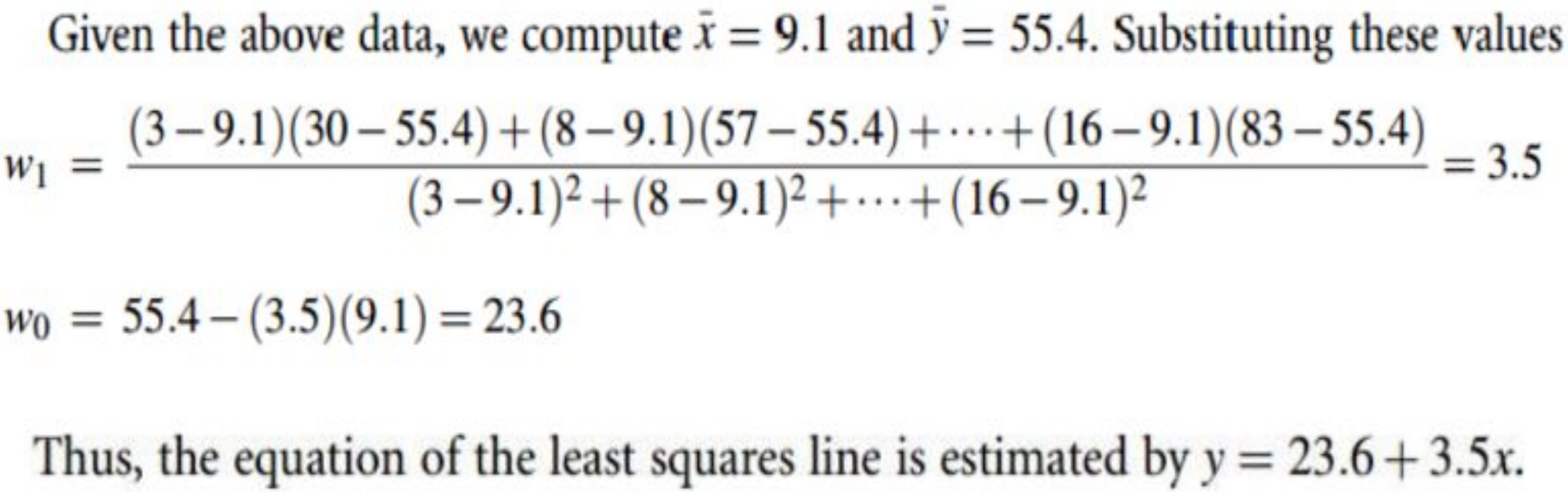
Linear Regression - Question



Salary data.

x *years experience*	y *salary (in \$1000s)*
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Form the linear regression equation for the given dataset



Given the above data, we compute $\bar{x} = 9.1$ and $\bar{y} = 55.4$. Substituting these values

$$w_1 = \frac{(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\cdots+(16-9.1)(83-55.4)}{(3-9.1)^2+(8-9.1)^2+\cdots+(16-9.1)^2} = 3.5$$

$$w_0 = 55.4-(3.5)(9.1) = 23.6$$

Thus, the equation of the least squares line is estimated by $y = 23.6+3.5x$.

Linear Regression - Types

upGrad

- Linear regression can be classified into two types: Simple Linear Regression and Multiple Linear Regression.
- Simple Linear Regression involves using one independent variable to model the relationship between that variable and a dependent variable.
- Multiple Linear Regression involves using more than one independent variable to model the relationship with the dependent variable.

Linear Regression - Types

Aspect	Simple Linear Regression	Multiple Linear Regression
Definition	A statistical method for finding a linear relationship between two variables.	A statistical method for finding a linear relationship between more than two variables.
Number of independent variables	One.	More than one.
Number of dependent variables	One.	One.
Equation	$y = mx + b$	$y = b + m_1x_1 + m_2x_2 + ... + m_nx_n$

Linear Regression - Question

Which of the following is an example of a categorical variable?

1. Age
2. Weight
3. Income
4. Gender

Linear Regression - Question

upGrad

Which of the following is an example of a categorical variable?

1. Age
2. Weight
3. Income
4. Gender

Linear Regression - Question

Which of the following is not an assumption of linear regression?

1. Normality of the residuals
2. Linearity of the relationship between the predictor and outcome variables
3. Homogeneity of variance
4. Independence of observations

Linear Regression - Question

Which of the following is not an assumption of linear regression?

1. Normality of the residuals
2. Linearity of the relationship between the predictor and outcome variables
3. Homogeneity of variance
4. Independence of observations

Linear Regression - Assumption

- The relationship between the independent and dependent variables is linear.
- The residuals, or errors, are normally distributed with a mean of zero and a constant variance.
- The independent variables are not correlated with each other (i.e. they are not collinear).
- The residuals are independent of each other (i.e. they are not autocorrelated).
- The model includes all the relevant independent variables needed to accurately predict the dependent variable

Clustering

- Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.
- Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures

Clustering - Examples

- Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, Web search, biology, and security.
- In business intelligence, clustering can be used to organize a large number of customers into groups, where customers within a group share strong similar characteristics. This facilitates the development of business strategies for enhanced customer relationship management.

Clustering - Examples

- In image recognition, clustering can be used to discover clusters or “subclasses” in handwritten character recognition systems. Suppose we have a data set of handwritten digits, where each digit is labeled as either 1, 2, 3, and so on. Note that there can be a large variance in the way in which people write the same digit. Take the number 2, for example. Some people may write it with a small circle at the left bottom part, while some others may not. We can use clustering to determine subclasses for “2,” each of which represents a variation on the way in which 2 can be written. Using multiple models based on the subclasses can improve overall recognition accuracy

Clustering

- Clustering is sometimes called **automatic classification**. Again, a critical difference here is that clustering can automatically find the groupings. This is a distinct advantage of cluster analysis.
- Clustering is also called **data segmentation** in some applications because clustering partitions large data sets into groups according to their similarity.
- Clustering can also be used for **outlier detection**, where outliers may be more interesting than common cases.
- Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce. For example, exceptional cases in credit card transactions, such as very expensive and infrequent purchases, may be of interest as possible fraudulent activities.

Clustering - Question

Clustering is a form of ________

1. Supervised Learning
2. Unsupervised Learning
3. Learning from Observation
4. Learning from Experience

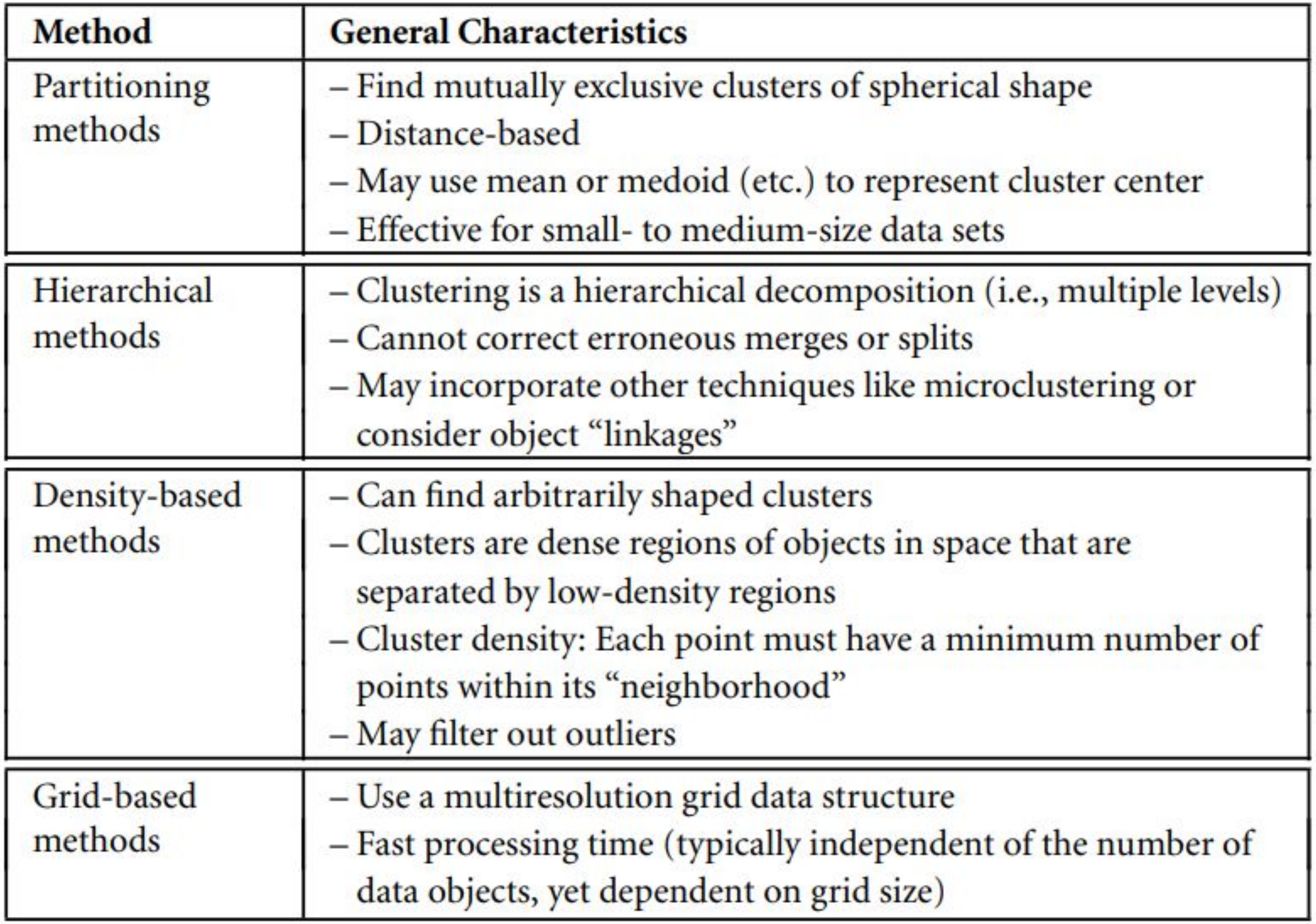
Clustering - Question

Clustering is a form of __________

1. Supervised Learning
2. Unsupervised Learning
3. Learning from Observation
4. Learning from Experience

Clustering - Methods

upGrad



Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Clustering - Partition Method - Kmeans - Question

Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are

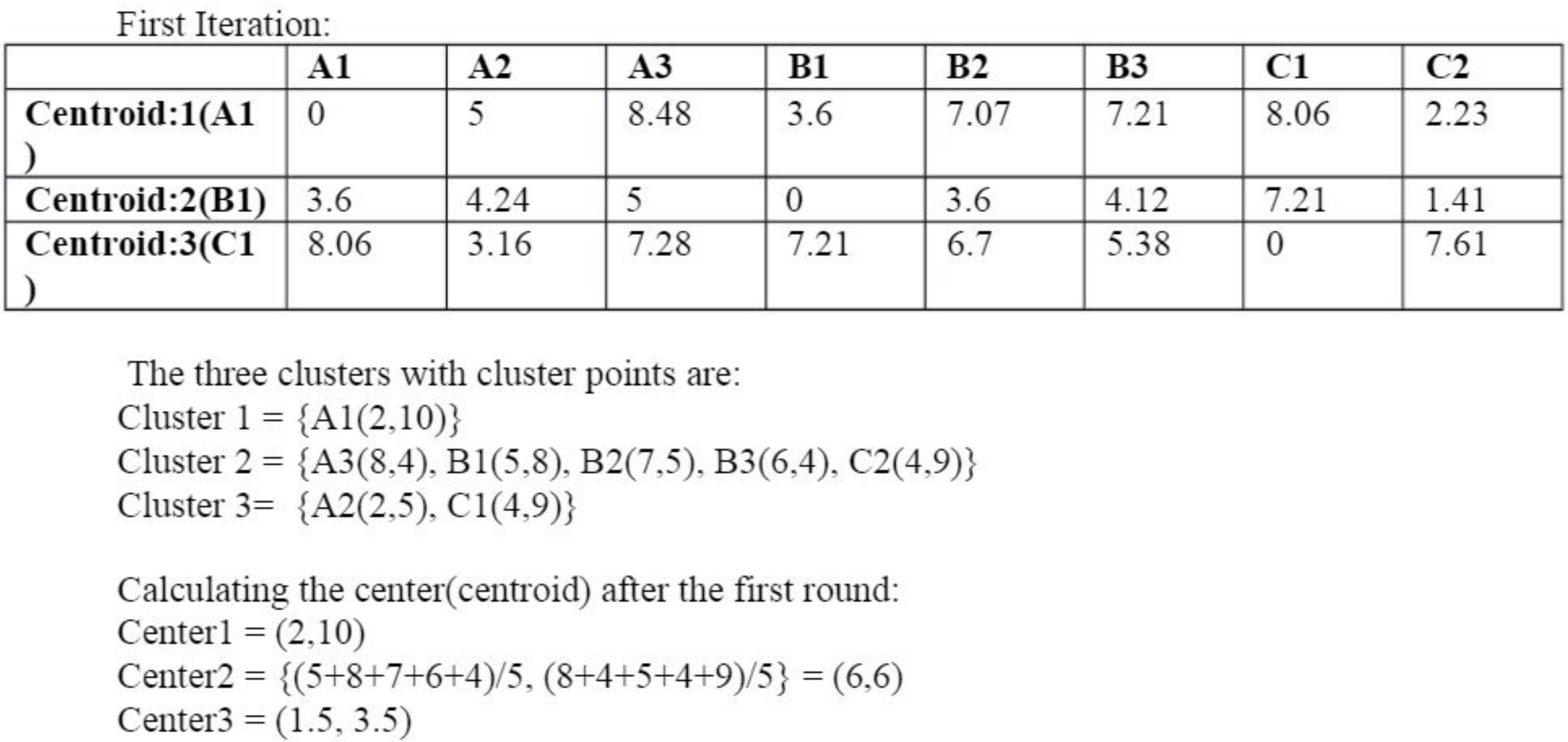
A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9).

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show only

(a) The three cluster centers after the first round of execution.

(b) The final three clusters.

Clustering - Partition Method - Kmeans - Question



First Iteration:

	A1	A2	A3	B1	B2	B3	C1	C2
Centroid:1(A1)	0	5	8.48	3.6	7.07	7.21	8.06	2.23
Centroid:2(B1)	3.6	4.24	5	0	3.6	4.12	7.21	1.41
Centroid:3(C1)	8.06	3.16	7.28	7.21	6.7	5.38	0	7.61

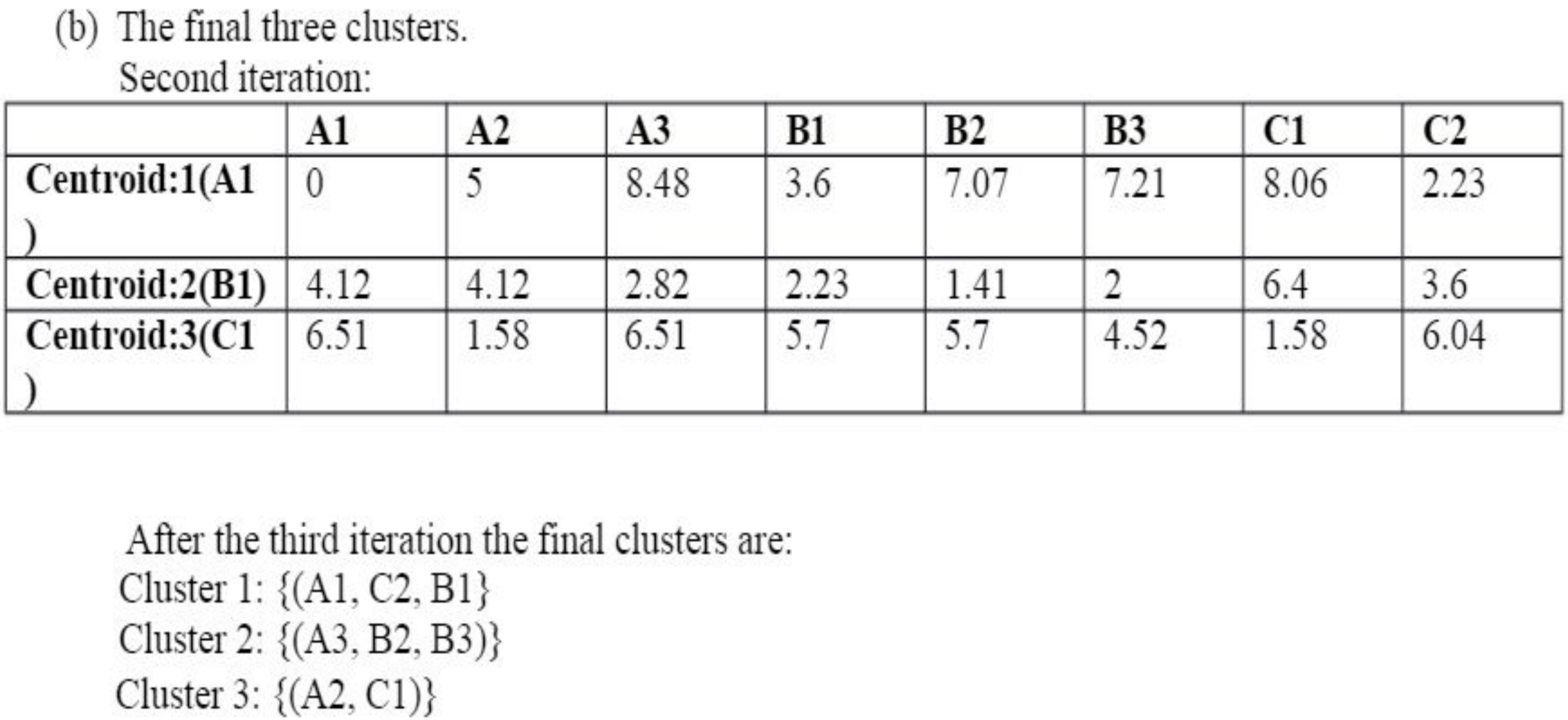
The three clusters with cluster points are:

Cluster 1 = {A1(2,10)}
Cluster 2 = {A3(8,4), B1(5,8), B2(7,5), B3(6,4), C2(4,9)}
Cluster 3= {A2(2,5), C1(4,9)}

Calculating the center(centroid) after the first round:

Center1 = (2,10)
Center2 = {(5+8+7+6+4)/5, (8+4+5+4+9)/5} = (6,6)
Center3 = (1.5, 3.5)

Clustering - Partition Method - Kmeans - Question



(b) The final three clusters.

Second iteration:

	A1	A2	A3	B1	B2	B3	C1	C2
Centroid:1(A1)	0	5	8.48	3.6	7.07	7.21	8.06	2.23
Centroid:2(B1)	4.12	4.12	2.82	2.23	1.41	2	6.4	3.6
Centroid:3(C1)	6.51	1.58	6.51	5.7	5.7	4.52	1.58	6.04

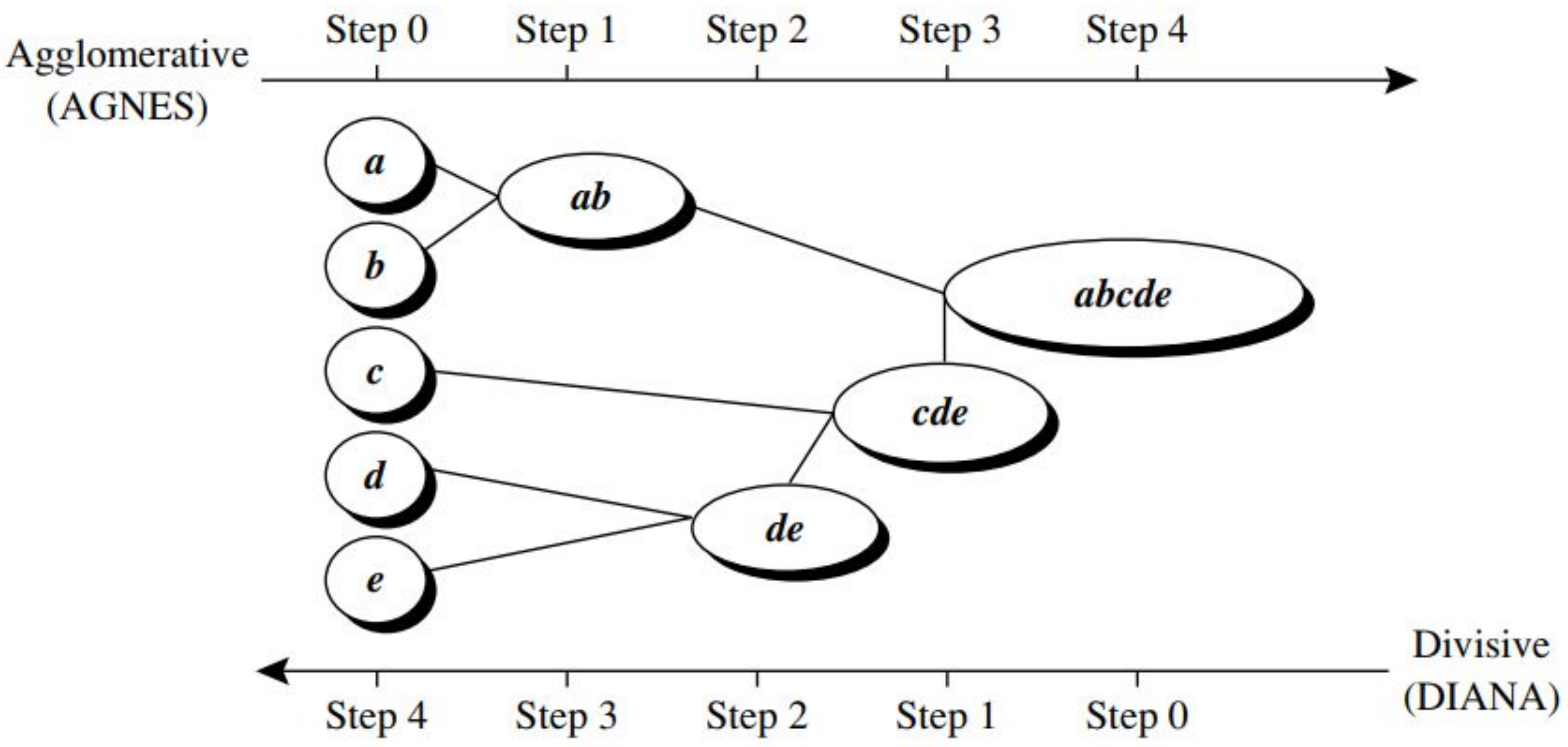
After the third iteration the final clusters are:

Cluster 1: {(A1, C2, B1}

Cluster 2: {(A3, B2, B3)}

Cluster 3: {(A2, C1)}

Clustering - Hierarchical Method



Clustering - Questions

Which of the following is finally produced by Hierarchical Clustering?

1. final estimate of cluster centroids
2. tree showing how close things are to each other
3. assignment of each point to clusters
4. all of the mentioned

Clustering - Questions

Which of the following is finally produced by Hierarchical Clustering?

1. final estimate of cluster centroids
2. tree showing how close things are to each other
3. assignment of each point to clusters
4. all of the mentioned

Clustering - Questions

Which of the following is required by K-means clustering?

1. defined distance metric
2. number of clusters
3. initial guess as to cluster centroids
4. all of the mentioned

Clustering - Questions

Which of the following is required by K-means clustering?

1. defined distance metric
2. number of clusters
3. initial guess as to cluster centroids
4. all of the mentioned

Clustering - Questions

Which of the following clustering requires merging approach?

1. Partitional
2. Hierarchical
3. Naive Bayes
4. None of the mentioned

Clustering - Questions

Which of the following clustering requires merging approach?

1. Partitional
2. [✓ Hierarchical]
3. Naive Bayes
4. None of the mentioned

THANK YOU