**Unit 5**

**Unsupervised Learning**

**Goal of unsupervised learning** - to discover the hidden patterns or structures of the data in which no target variable exists to perform either classification or regression methods.

**Unsupervised Learning methods**

- challenging, as the outcomes are subjective and no simple goal for the analysis (predicting the class or continuous variable). These methods are performed as part of exploratory data analysis.

- hard to assess the results, since there is no universally accepted mechanism for performing the validation of results.

**Unsupervised Learning Applications of are:**

* **Genomics:** Unsupervised learning applied to understand genomic-wide biological insights from DNA to better understand diseases and peoples.
* **Search engine:** Search engines might choose which search results to display to a particular individual based on the click histories of other similar users.
* **Knowledge extraction:** To extract the taxonomies of concepts from raw text to generate the knowledge graph to create the semantic structures in the field of NLP.
* **Segmentation of customers:** In the banking industry, unsupervised learning like clustering is applied to group similar customers, and based on those segments, marketing departments design their contact strategies. For example, older, low-risk customers will be targeted with fixed deposit products and high-risk, younger customers will be targeted with credit cards or mutual funds, and so on.
* **Social network analysis:** To identify the cohesive groups of people in social networks who are more connected with each other and have similar characteristics in common.

**Unsupervised Learning techniques:**

* K-means clustering
* Principal component analysis
* Singular value decomposition

**K-means clustering**

Clustering is the task of grouping observations in such a way that members of the same cluster are more similar to each other, and members of different clusters are very different from each other.

Clustering is used to explore a dataset to either identify the underlying patterns in it or to create a group of characteristics.

**Example:** Social networks can be clustered to identify communities and to suggest missing connections between people.

* In anti-money laundering measures, suspicious activities and individuals can be identified using anomaly detection
* In biology, clustering is used to find groups of genes with similar expression patterns
* In marketing analytics, clustering is used to find segments of similar customers so that different marketing strategies can be applied to different customer segments accordingly

**k-means clustering algorithm** - an iterative process of moving the centers of clusters or centroids to the mean position of their constituent points, and reassigning instances to their closest clusters iteratively until there is no significant change in the number of cluster centers possible or number of iterations reached.

The cost function of k-means is determined by the Euclidean distance (square-norm) between the observations belonging to that cluster with its respective centroid value.

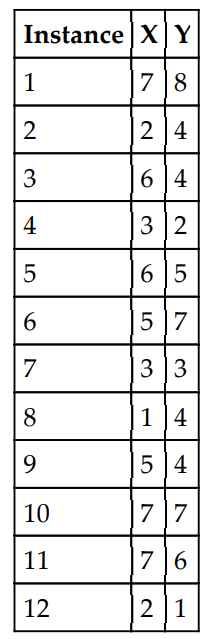
**Equation** - if there is only one cluster (k=1), then the distances between all the observations are compared with its single mean. If number of clusters increases to 2 (k= 2), then two-means are calculated and a few of the observations are assigned to cluster 1 and other observations are assigned to cluster two-based on proximity. Subsequently, distances are calculated in cost functions by applying the same distance measure, but separately to their cluster centers:

A picture containing text, clock

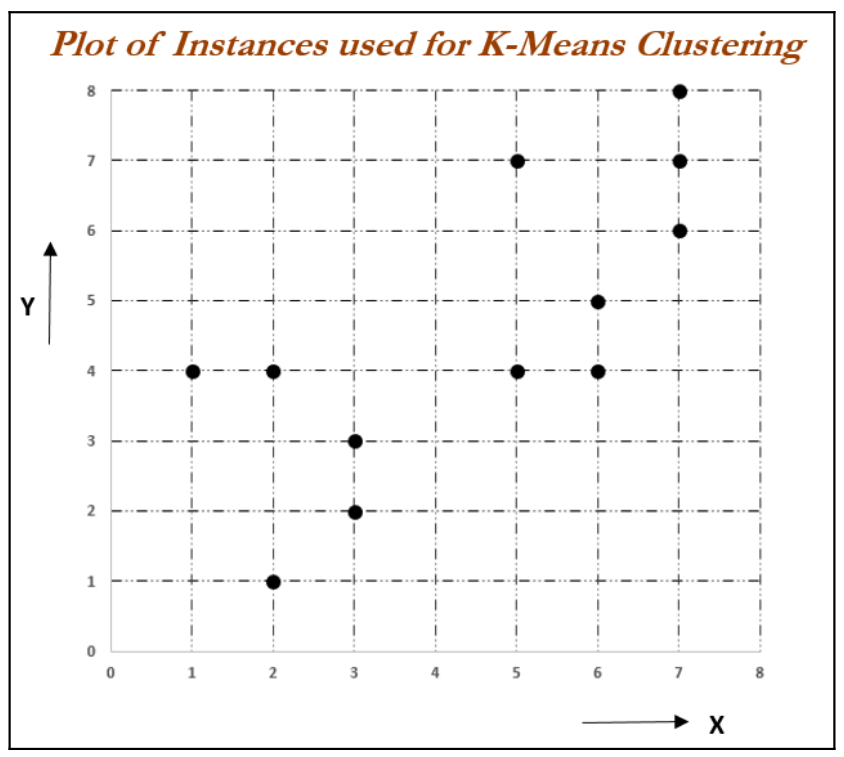
Description automatically generated

**K-means working methodology from first principles**

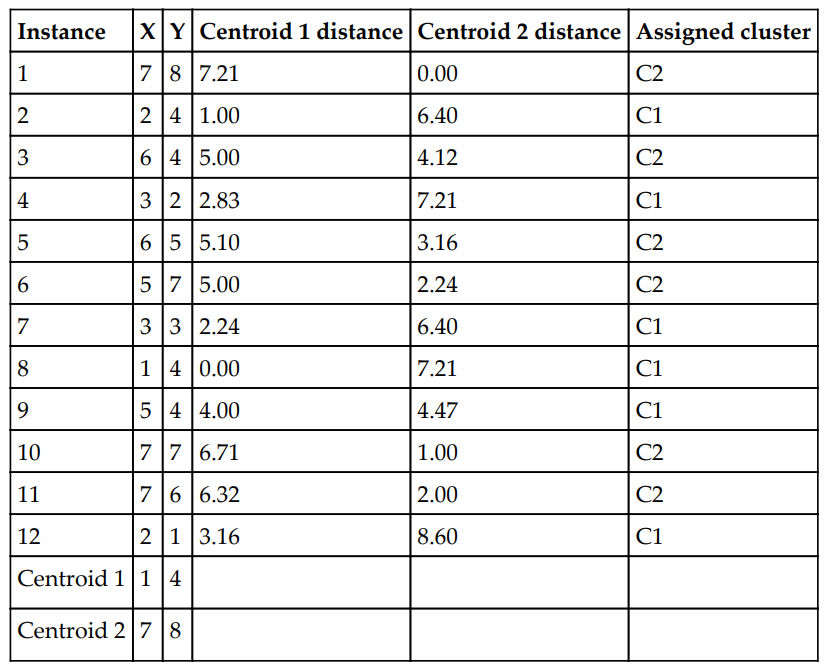
The k-means working methodology is illustrated in the following example in which 12 instances are considered with their X and Y values. The task is to determine the optimal clusters out of the data.



After plotting the data points on a 2D chart, roughly two clusters are possible, where below-left is the first cluster and the top-right is another cluster, but in many practical cases there would be so many variables (or dimensions) that, we cannot simply visualize them. Hence, we need a mathematical and algorithmic way to solve these types of problems.



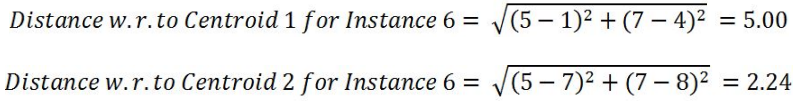
**Iteration 1:** Assume two centers from two instances out of all the 12 instances. Here, we have chosen instance 1 (X = 7, Y = 8) and instance 8 (X = 1, Y = 4), as they seem to be at both extremes. For each instance, we will calculate its Euclidean distances with respect to both centroids and assign it to the nearest cluster center.



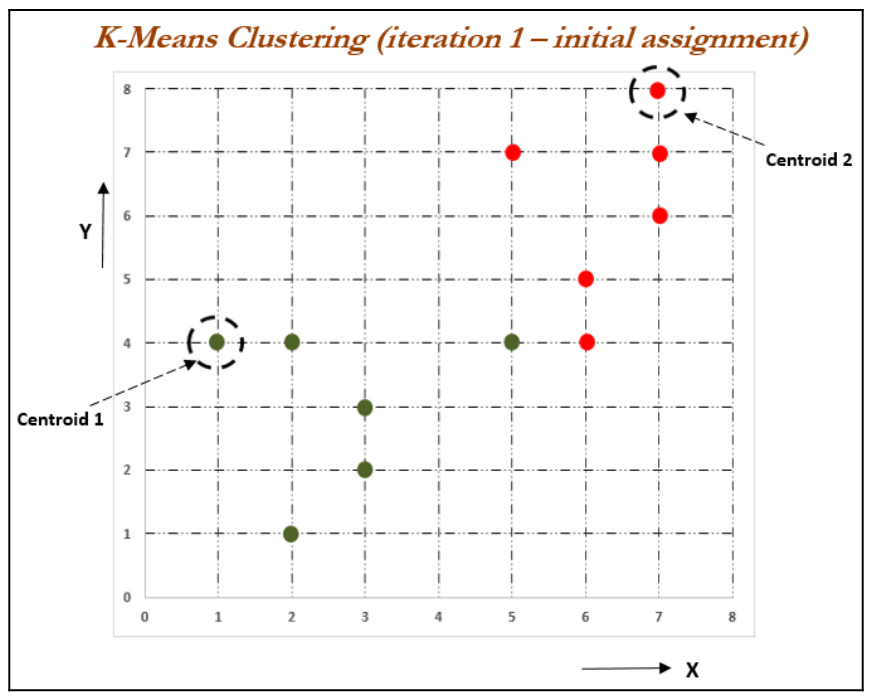
The Euclidean distance between two points A (X1, Y1) and B (X2, Y2) is shown as follows:



Centroid distance calculations are performed by taking Euclidean distances. A sample calculation has been shown as follows. For instance, six with respect to both centroids (centroid 1 and centroid 2).

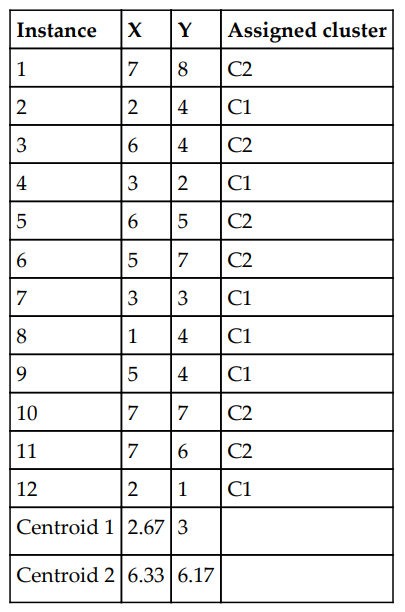


The following chart describes the assignment of instances to both centroids, which was shown in the preceding table format:

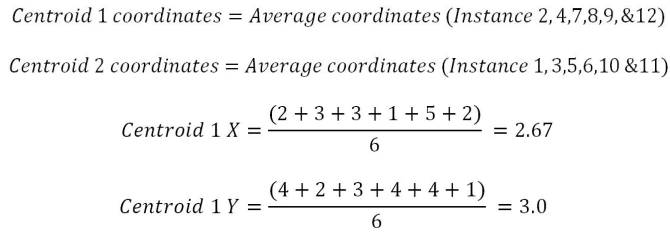


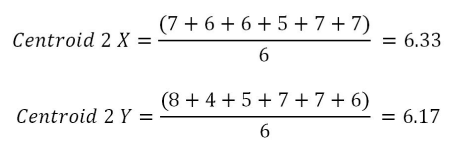
If we observe the preceding chart, all the instances seem to be assigned appropriately apart from instance 9 (X =5, Y = 4). However, in later stages, it should be assigned appropriately.

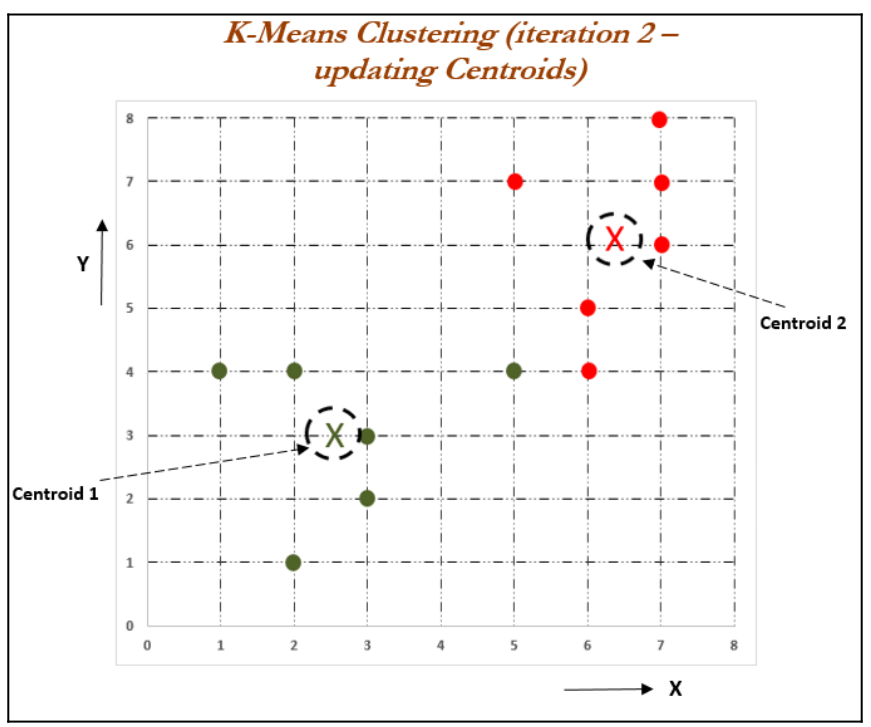
**Iteration 2:** In this iteration, new centroids are calculated from the assigned instances for that cluster or centroid. New centroids are calculated based on the simple average of the assigned points.



Sample calculations for centroids 1 and 2 are shown as follows. A similar methodology will be applied on all subsequent iterations as well:

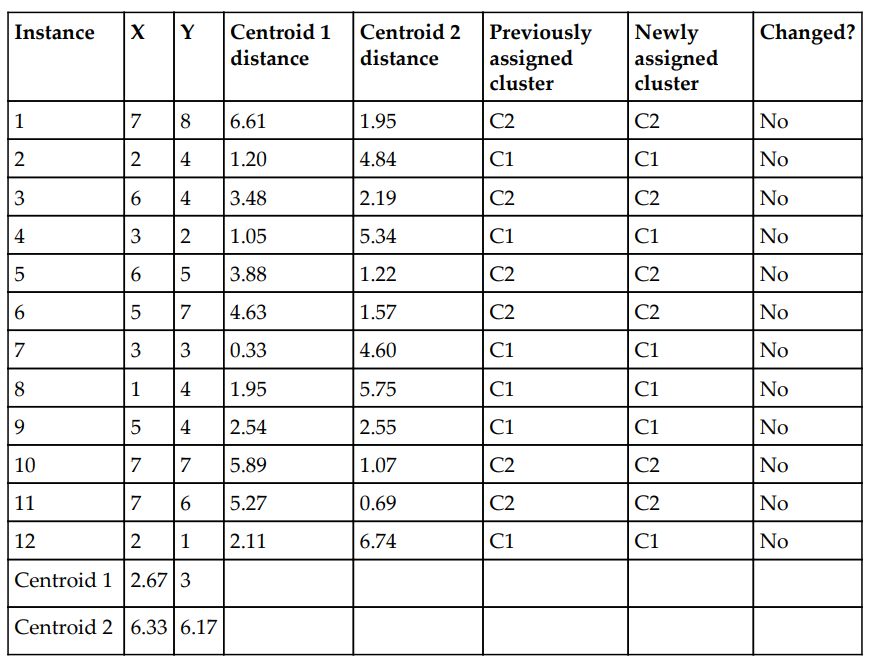






After updating the centroids, we need to reassign the instances to the nearest centroids, which we will be performing in iteration 3.

**Iteration 3:** In this iteration, new assignments are calculated based on the Euclidean distance between instances and new centroids. In the event of any changes, new centroids will be calculated iteratively until no changes in assignments are possible or the number of iterations is reached. The following table describes the distance measures between new centroids and all the instances:



It seems that there are no changes registered. Hence, we can say that the solution is converged. One important thing to note here is that all the instances are very clearly classified well, apart from instance 9 (X = 5, Y = 4). Based on instinct, it seems like it should be assigned to centroid 2, but after careful calculation, that instance is more proximate to cluster 1 than cluster 2. However, the difference in distance is low (2.54 with centroid 1 and 2.55 with centroid 2).

**Optimal number of clusters and cluster evaluation**

Selecting number of clusters

**The elbow method**

* Used to determine the optimal number of clusters in k-means clustering.
* The elbow method plots the value of the cost function produced by different values of k.
* If k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids.
* The improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

Chart, line chart

Description automatically generated

**Evaluation of clusters with silhouette coefficient:**

* It is a measure of the compactness and separation of the clusters.
* Higher values represent a better quality of cluster.
* The silhouette coefficient is higher for compact clusters that are well separated and lower for overlapping clusters.
* Silhouette coefficient values do change from -1 to +1, and the higher the value is, the better.
* The silhouette coefficient is calculated per instance. For a set of instances, it is calculated as the mean of the individual sample's scores.

Text

Description automatically generated with low confidence

* a is the mean distance between the instances in the cluster, b is the mean distance between the instance and the instances in the next closest cluster.

**Principal component analysis - PCA**

* PCA is the dimensionality reduction technique.
* PCA reduces the dimensions of a dataset by projecting the data onto a lower-dimensional subspace.
* Example: a 2D dataset could be reduced by projecting the points onto a line. Each instance in the dataset would then be represented by a single value, rather than a pair of values.
* In a similar way, a 3D dataset could be reduced to two dimensions by projecting variables onto a plane.

**PCA utilities:**

* Mitigate the course of dimensionality
* Compress the data while minimizing the information lost at the same time
* Principal components will be further utilized in the next stage of supervised learning, in random forest, boosting, and so on.
* Understanding the structure of data with hundreds of dimensions can be difficult, hence, by reducing the dimensions to 2D or 3D, observations can be visualized easily

The left-hand side of the diagram depicts the top view, front view, and side view of the component. On the right-hand side, an isometric view has been drawn, in which one single image has been used to visualize how the component looks. So, one can imagine that the left-hand images are the actual variables, and the right-hand side is the first principal component, in which most variance has been captured.

Finally, three images have been replaced by a single image by rotating the axis of direction. We replicate the same technique in PCA analysis.

Diagram

Description automatically generated

Principal component working methodology is explained in the following example, in which actual data has been shown in a 2D space, in which X and Y axis are used to plot the data. Principal components are the ones in which maximum variation of the data is captured.

Chart, scatter chart

Description automatically generated

The following diagram illustrates how it looks after fitting the principal components. The first principal component covers the maximum variance in the data and the second principal component is orthogonal to the first principal component, and all principal components are orthogonal to each other.

We can represent whole data with the first principal component itself. That is how it is advantageous to represent the data with fewer dimensions, to save space and also to grab maximum variance in the data, which can be utilized for supervised learning in the next stage.

Chart, scatter chart

Description automatically generated

Basics on eigenvectors and eigenvalues is necessary when studying PCAs.

Eigenvectors are the axes (directions) along which a linear transformation acts simply by stretching/compressing and/or flipping.

eigenvalues give the factors by which the compression occurs. An eigenvector of a linear transformation is a nonzero vector whose direction does not change when that linear transformation is applied to it.

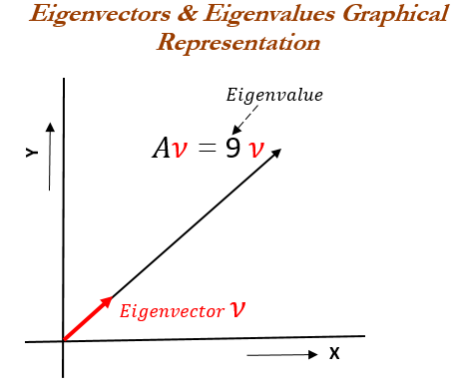
A is a linear transformation from a vector space and  is a nonzero vector, then eigen vector of A if is a scalar multiple of . The condition can be written as the following equation:



In the above equation,  is an eigenvector, A is a square matrix, and λ is a scalar called an eigenvalue.

The direction of an eigenvector remains the same after it has been transformed by A; only its magnitude has changed, as indicated by the eigenvalue, That is, multiplying a matrix by one of its eigenvectors is equal to scaling the eigenvector, which is a compact representation of the original matrix.

The following graph describes eigenvectors and eigenvalues in a graphical representation in a 2D space:



The following example describes how to calculate eigenvectors and eigenvalues from the square matrix. Note: eigenvectors and eigenvalues can be calculated only for square matrices.

Diagram, text

Description automatically generated with medium confidence

The product of A and any eigenvector of A must be equal to the eigenvector multiplied by the magnitude of eigenvalue:

A picture containing text, clock

Description automatically generated

A characteristic equation states that the determinant of the matrix, that is the difference between the data matrix and the product of the identity matrix and an eigenvalue is 0.

A picture containing text, clock

Description automatically generated

(2-λ)(-6-λ) – (-16)

Both eigenvalues for the preceding matrix are equal to -2. We can use eigenvalues to substitute for eigenvectors in an equation:

Text, letter

Description automatically generated

Substituting the value of eigenvalue in the preceding equation, we will obtain the following formula:

Diagram

Description automatically generated

The preceding equation can be rewritten as a system of equations, as follows:

Text

Description automatically generated with low confidence

This equation indicates it can have multiple solutions of eigenvectors we can substitute with any values which hold the preceding equation for verification of equation. Here, we have used the vector [1 1] for verification, which seems to be proved.

A picture containing text, clock

Description automatically generated

PCA needs unit eigenvectors to be used in calculations, hence we need to divide the same with the norm or we need to normalize the eigenvector. The 2-norm equation is shown as follows:

Diagram, schematic

Description automatically generated

The norm of the output vector is calculated as follows:

A picture containing text, clock

Description automatically generated

The unit eigenvector is shown as follows:

Text

Description automatically generated with medium confidence

**Singular value decomposition - SVD**

Many implementations of PCA use singular value decomposition to calculate eigenvectors and eigenvalues. SVD is given by the following equation:

Text, letter

Description automatically generated

Columns of U are called left singular vectors of the data matrix, the columns of V are its right singular vectors, and the diagonal entries of are its singular values. Left singular vectors are the eigenvectors of the covariance matrix and the diagonal element of are the square roots of the eigenvalues of the covariance matrix.

Before proceeding with SVD, it would be advisable to understand a few advantages and

important points about SVD:

* SVD can be applied even on rectangular matrices; whereas, eigenvalues are defined only for square matrices. The equivalent of eigenvalues obtained through the SVD method are called singular values, and vectors obtained equivalent to eigenvectors are known as singular vectors. However, as they are rectangular in nature, we need to have left singular vectors and right singular vectors respectively for their dimensions.
* If a matrix A has a matrix of eigenvectors P that is not invertible, then A does not have an eigen decomposition. However, if A is m x n real matrix with m > n, then

A can be written using a singular value decomposition.

Text

Description automatically generated

Then computation of singular values and singular vectors is done with the following set of equations:



In the first stage, singular values/eigenvalues are calculated with the equation. Once we obtain the singular/eigenvalues, we will substitute to determine the V or right

singular/eigen vectors:

Once we obtain the right singular vectors and diagonal values, we will substitute to obtain the left singular vectors U using the equation mentioned as follows:



In this way, we will calculate the singular value decompositions of the original system of

equations matrix.

Unit 1

Statistical model, Machine learning model, Linear regression, Gradient descent, Train, Test, Validation, Grid search

Unit 2

Linear regression, Ridge and Lasso regression, Random Forest, Logistic regression, Grid search on RF

Unit 3

KNN, Naïve Bayes

Unit 4

SVM, Kernel functions, ANN, Forward and Backward propagation, Optimization of NNs – Stochastic gradient descent, Deep learning

Unit 5

K-means clustering, elbow method, PCA, SVD