Vivek Muskan
231FA04G03
Section- 12 ('L')

"Machine Learning"

"Spectral Clustering"

Spectral Clustering is a powerful unsupervised learning algorithm used to group data points into clusters based on their similarity.

Unlike traditional methods like K-means (which work well only for circular clusters), spectral clustering can handle complex shapes and non-linear boundaries.

It is based on concepts from graph theory - the data is represented as a graph, where each data point is a node and the edges show how similar two points are.

Why Spectral Clustering ?

Imagine, we have data points that form two 'half-moon shapes'.

Traditional clustering methods like K-means will fail because the clusters are not clearly separated by straight lines.

Spectral clustering, however can separate them easily

because it looks at the connections between points (not just distances).

So, instead of directly grouping data, spectral clustering:

- Builds a graph from the data.
- Finds the important structure of the graph using eigenvalues and eigenvectors.
- Uses that structure to form the final clusters.

✳ Steps in Spectral Clustering

Step 1. Create a Similarity Matrix

- Compute how similar each pair of points is.
- Often done using a Gaussian kernel:

$$\text{Similarity}(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

- This forms a matrix 'W', where each entry W_{ij} shows similarity between points i & j.

Step 2. Build the Graph Laplacian

- Compute the degree matrix (D) — a diagonal matrix where each diagonal element is the sum of similarities for that node.
- Then compute the Laplacian matrix:

$$L = D - W$$

Step 3. Find Eigenvectors.

- Compute the eigenvalues and eigenvectors of the Laplacian matrix 'L'.
- Select the 'K' smallest eigenvectors (Where, K = no. of clusters).
- These eigenvectors form a new Feature space.

Step 4. Apply K-Means

- Run the K-means algorithm on the new feature space.
- The output clusters correspond to the final special clusters.

Example Dataset :-

"Iris Dataset"

↳ The Iris dataset is a classic dataset used for classification. It contains measurements of flowers from the three different species of Iris:-

- Iris - Sentosa
- Iris - versicolor
- Iris - virginica

So, in this dataset, spectral clustering works very well. This is because of the following reasons:-

- Iris dataset clusters are not perfectly spherical in all feature dimensions.
- Traditional K-means may fail to seperate classes perfectly.
- Spectral clustering uses graph-based similarities and can capture complex cluster shapes.

So, this is the greatest advantage of Spectral clustering over earlier clustering methods. These seperate the classes well for image datasets.