29. Comparative Analysis of Clustering Approaches: K-means with Piecewise Linear Decision Boundaries vs. Gaussian Mixture Models for Patient Segmentation

**Abstract**

The segmentation of patient populations based on critical health metrics is fundamental to precision medicine. This study compares two clustering methodologies—K-means with piecewise linear decision boundaries and Gaussian Mixture Models (GMM) with common covariance and Bayes decision boundaries—to classify patients by blood pressure (mmHg) and cholesterol levels (mg/dL). The analysis reveals that while K-means provides a straightforward and efficient clustering with linear boundaries, GMM offers a more flexible partitioning with curved decision boundaries. The results underscore the importance of selecting an appropriate clustering method based on the specific characteristics of the data distribution and the clinical objectives of the segmentation.

**Introduction**

In healthcare, patient segmentation is essential for identifying groups at varying risk levels and tailoring treatments accordingly. Blood pressure and cholesterol are key metrics for assessing cardiovascular risk. Clustering techniques such as K-means and Gaussian Mixture Models (GMM) can help identify distinct patient groups based on these parameters. However, these methods differ significantly in their assumptions and performance.

K-means clustering is a popular choice due to its simplicity and computational efficiency. It partitions the data into clusters by minimizing within-cluster variance and uses linear decision boundaries. However, this method assumes that clusters are spherical and equally sized, which may not always be valid. In contrast, GMM allows for more flexibility by assuming that the data is generated from a mixture of several Gaussian distributions. This study compares these two approaches to highlight their strengths and limitations in patient segmentation.

**Methods**

The dataset comprises patient records with two key variables: blood pressure (mmHg) and cholesterol levels (mg/dL). We employed two clustering techniques:

1. **K-means Clustering with Piecewise Linear Decision Boundaries:** The K-means algorithm was applied to divide the patients into five clusters based on their blood pressure and cholesterol levels. The piecewise linear decision boundaries (blue dashed lines) were drawn to illustrate the separation between clusters, assuming linear divisions between groups.
2. **Gaussian Mixture Models (GMM) with Common Covariance and Bayes Decision Boundaries:** GMM clustering was also applied, with the assumption that each cluster follows a Gaussian distribution. The GMM model was implemented with a common covariance structure, and Bayes decision boundaries (purple dotted lines) were derived to show the probabilistic separations between clusters.

**Results**

The plots provide a visual comparison of the clustering results from the two approaches:

1. **K-means Clustering with Piecewise Linear Decision Boundaries:**
   * The five clusters (indicated by different colors) show clear linear separations.
   * Decision boundaries are straight lines, illustrating the simplicity of the K-means method in defining clusters.
   * While efficient, the linear boundaries may not capture more nuanced relationships between the variables, potentially oversimplifying the cluster definitions.
2. **Gaussian Mixture Models (GMM) with Common Covariance and Bayes Decision Boundaries:**
   * The clusters generated by GMM are represented with more flexible, curved boundaries (purple dotted lines), reflecting the probabilistic nature of the model.
   * The boundaries are smoother and adapt better to variations within the data, suggesting that GMM can capture more complex relationships between blood pressure and cholesterol levels.
   * The GMM clustering shows how clusters overlap more naturally, accounting for the inherent uncertainty and variability in patient health metrics.

**Discussion**

The comparison of K-means and GMM for patient segmentation based on blood pressure and cholesterol levels provides valuable insights into the strengths and limitations of each approach:

* **K-means Clustering:** This method is advantageous for datasets with clear, distinct groups where linear separations are appropriate. Its simplicity and speed make it a practical choice for large datasets. However, K-means may fall short when the data exhibits more complex patterns that require curved or non-linear boundaries.
* **Gaussian Mixture Models (GMM):** GMM offers greater flexibility by assuming that each cluster follows a Gaussian distribution. This allows for capturing overlapping clusters and adapting to various data shapes. The curved Bayes decision boundaries provide a more refined view of the cluster separations, especially in datasets where patient groups are not well-separated by linear boundaries.

**Conclusion**

Both K-means clustering and Gaussian Mixture Models have their place in patient segmentation. K-means provides a quick and efficient solution with linear decision boundaries, while GMM offers a more flexible approach with curved boundaries that better reflect the underlying data distributions. The choice of clustering method should depend on the specific data characteristics and the clinical goals of the segmentation.

Future research should explore the integration of additional clinical variables and the application of these clustering methods in a longitudinal context to assess their effectiveness in predicting patient outcomes.

**References**

* Cite relevant literature on K-means clustering, Gaussian Mixture Models, and their applications in medical data analysis.

A group of colorful lines

Description automatically generated with medium confidence