**AMA 546 Project Speech**

**Page 11**

We selected three evaluation metrics, Silhouette Score(s), Calinski-Harabasz Index(CHI), Davies-Bouldin index(DBI). The calculation formulas of the three metrics are as shown in the slide.

**Page 12**

……

**Page 13**

Gaussian Mixture Model (GMM) is a clustering method based on probability density function.

When clustering, GMM assigns data points to the cluster with the highest probability, instead of hard assigning data points to a certain cluster like K-Means.

GMM can be applied to multiple fields, such as image recognition, speech processing, risk assessment, etc.

**Page 14**

Affinity Propagation (AP) is a clustering algorithm used to group data points into clusters with similar characteristics.

It is an algorithm based on graph theory that determines the center point of a cluster by calculating the similarity between data points.

AP clustering don’t need to specify the number of final clustering families, and is not sensitive to the initial value of the data. Compared with the k-centers clustering method, the squared error of the result is smaller.

**Page 15**

Next, we will show the process of clustering this dataset by the above model.

**Page 16**

In K-means, we using the Elbow Method to Determine Optimal Clustering Number k. The range of k is from 1 to 11.

**Page 17**

Then, we utilized t-SNE to reduce the dimensions to a 2D space for visualization. The evaluation metrics can be seen on the right side of the silde.

**Page 18**

In Gaussian Mixture Model, We first need to determine the k value, The range of k is from 2 to 11. We notice that, when k=2, the Silhouette Score and CHI is higher while DBI is lower. So k=2 is what we want.

**Page 19**

Then, we try to adjust covariance matrix, the Silhouette Score are shown on the left side table. It’s clearly that ‘full’ type of covariance matrix is better than others.

**Page 20**

On this basis, we calculated other evaluation metrics.

**Page 21-22**

We visualize the clustering results and reflect the relationship between different categories through scatter plot, histogram, pie chart and pairplot.

**Page 23**

The experiment of Affinity Propagation Model is easier. We firstly determine the parameters to find the AP.label, then calculated all evaluation metrics.

**Page 24**

Finally, based on the above experimental results, we make the following comments.

**Page 25**

Firstly, we need to compare the performance of different models in this clustering task. There are three points of analysis.

1. The clustering results of different models on the same data are significantly different.
2. The Gaussian Mixture Model performs best among the three evaluation metrics.
3. In my view, there is no reason to set 224 clusters in this dataset, which is very irrational, and metrics have very bad scores either. So, the Affinity Propagation Model isn't useful for this cluster task.

**Page 26**

Taking the clustering result of K-means as an example, we further analyze the clustering results and draw the conclusion as follows:

1. The features of cluster 0 are low balance, frequent purchases and minimal cash advances.
2. The features of cluster 1 are high balances, low purchase frequency and high cash advance usage.
3. The features of cluster 2 are high balances, active purchasing and good credit.

**Page 27**

This is all the project content of our team, thank you for listening!