

K Modes

AIM

This project applied the K-Modes clustering algorithm to the categorical Titanic dataset (**Class**, **Sex**, **Age**, **Survived**) to uncover natural groupings. Silhouette analysis using Gower distance helped determine the optimal number of clusters.

EXPERIMENTAL SETUP

The dataset used is **Titanic**, converted into individual passenger records using the **Freq** column. The goal is to perform unsupervised clustering using categorical attributes.

LIBRARIES REQUIRED

- **klaR** – for the K-Modes algorithm
- **cluster** – for silhouette analysis
- **ggplot2** – for visualization
- **datasets** – for the Titanic dataset

STEPS

1. Load necessary libraries
2. Load and preprocess Titanic dataset
3. Create row-wise data from frequency-based table
4. Implement hyperparameter tuning over **k**
5. Calculate Gower distance and silhouette scores
6. Determine optimal number of clusters
7. Apply K-Modes with optimal **k**
8. Print cluster mode information
9. Visualize clusters by category (e.g., **Class**)

CODE SNIPPET

```
data("Titanic")
titanic_df <- as.data.frame(Titanic)

# Expand data so each row represents one passenger
titanic_cluster_df <- data.frame(
  Class = factor(rep(titanic_df$Class, titanic_df$Freq)),
  Sex = factor(rep(titanic_df$Sex, titanic_df$Freq)),
  Age = factor(rep(titanic_df$Age, titanic_df$Freq)),
  Survived = factor(rep(titanic_df$Survived, titanic_df$Freq))
)
set.seed(42)
k_values <- 2:10
silhouette_scores <- numeric(length(k_values))

for (i in seq_along(k_values)) {
  k <- k_values[i]
  kmodes_result <- kmodes(titanic_cluster_df, modes = k, iter.max = 100)
  cluster_labels <- kmodes_result$cluster

  dissimilarity <- daisy(titanic_cluster_df, metric = "gower")
  ss <- silhouette(cluster_labels, dissimilarity)
  silhouette_scores[i] <- mean(ss[, "sil_width"])
}

# Determine optimal number of clusters
optimal_k <- k_values[which.max(silhouette_scores)]
kmodes_final <- kmodes(titanic_cluster_df, modes = optimal_k, iter.max = 100)
```

```
cat("Optimal number of clusters:", optimal_k, "\n")
```

```
## Optimal number of clusters: 10
```

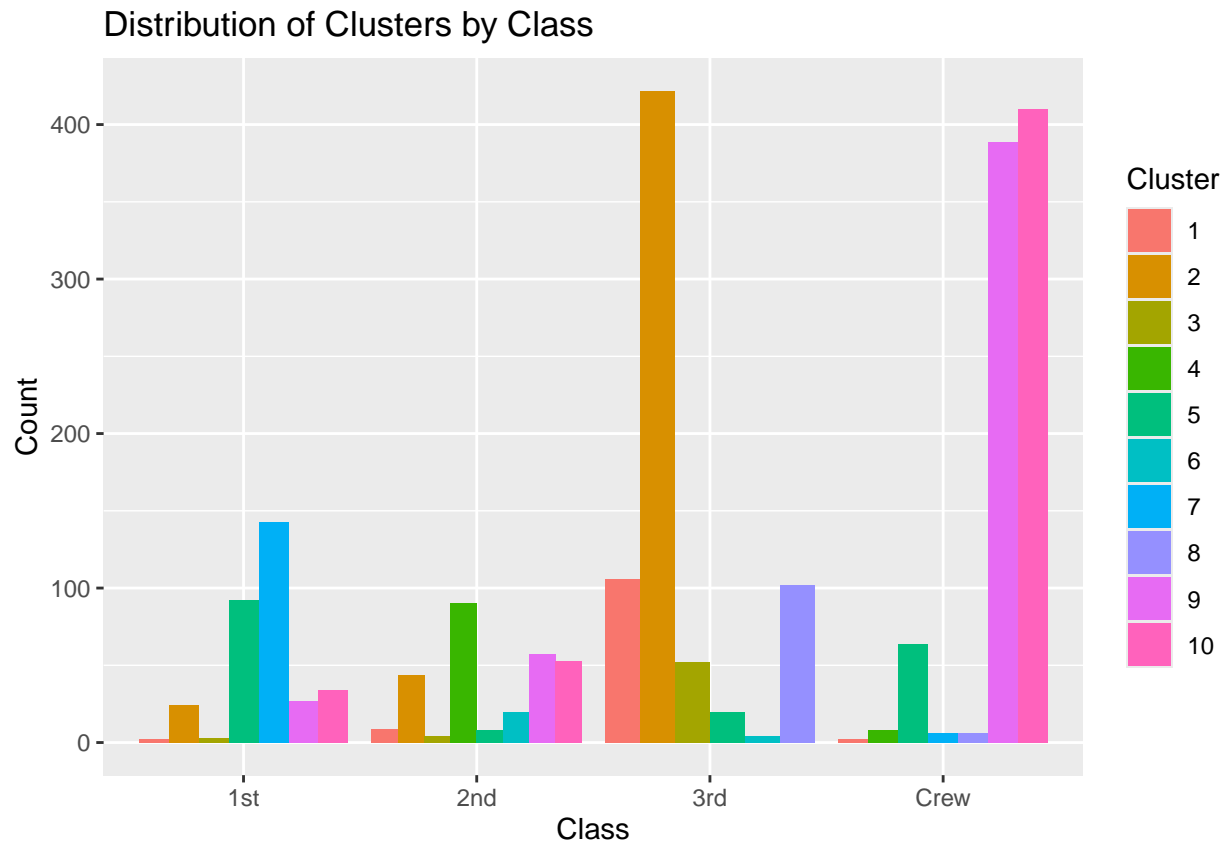
```
cat("Cluster mode information:\n")
```

```
## Cluster mode information:
```

```
print(kmodes_final$modes)
```

```
##      Class    Sex   Age Survived
## 1     3rd Female Adult         No
## 2     3rd  Male Adult         No
## 3     3rd  Male Child         Yes
## 4     2nd Female Adult         Yes
## 5     1st  Male Adult         Yes
## 6     2nd Female Child         Yes
## 7     1st Female Adult         Yes
## 8     3rd Female Adult         Yes
## 9     Crew  Male Adult         No
## 10    Crew  Male Adult         No
```

```
ggplot(titanic_cluster_df, aes(x = Class, fill = factor(kmodes_final$cluster))) +
  geom_bar(position = "dodge") +
  labs(
    title = "Distribution of Clusters by Class",
    x = "Class",
    y = "Count",
    fill = "Cluster"
  )
)
```



CONCLUSION

The analysis revealed distinct clusters, such as third-class male children with low survival rates, first-class male adults, and crew members with high survival.

These patterns align with historical data and demonstrate how **K-Modes** can uncover meaningful insights from categorical variables in unsupervised learning.