

ISI

Lab 10

Daniel Martínez Sánchez

15th May, 2023

Problem definition, dataset description and data preparation

Dataset loading

1.3.1. Based on the previous result, which two features have (on average) the smallest values and the largest values?

Feature with smallest value on average (1.3735933983495874): FlightTrans

Feature with highest value on average (73601.32758189547): Balance

1.3.2. Why is it important to normalize the data before clustering?

Normalizing the data before clustering is crucial because it ensures that all features are on a similar scale, allowing for fair comparisons and eliminating bias. It reduces the impact of outliers, enhances convergence speed, and facilitates the interpretation and comparison of clusters.

Data preprocessing

1.4.1. What are the mean and standard deviation of the features in the standardized dataset?

Mean of the standardized dataset:

- 0: 5.756558e-16
- 1: -7.588884e-17
- 2: -5.515024e-16
- 3: -2.877696e-15
- 4: 5.781822e-16
- 5: -1.206323e-15
- 6: 1.406727e-15

Standard deviation of the standardized dataset:

- 0: 1.000125
- 1: 1.000125
- 2: 1.000125
- 3: 1.000125

- 4: 1.000125
- 5: 1.000125
- 6: 1.000125

1.4.2. Based on the normalized dataset descriptive stats, which two features have (on average) the smallest values and the largest values?

Feature with smallest value on average ($-2.877695858827106e-15$): 3

Feature with highest value on average ($1.4067266981184467e-15$): 6

Building and Using Unsupervised Clustering Models

Building and using the clustering model

Cluster with highest average value for Balance, QualMiles, BonusMiles, BonusTrans, FlightMiles, FlightTrans, DaysSinceEnroll:

- Balance: 1
- QualMiles: 5
- BonusMiles: 4
- BonusTrans: 1
- FlightMiles: 1
- FlightTrans: 1
- DaysSinceEnroll: 2

Cluster with lowest average value for Balance, QualMiles, BonusMiles, BonusTrans, FlightMiles, FlightTrans, DaysSinceEnroll:

- Balance: 3
- QualMiles: 3
- BonusMiles: 3
- BonusTrans: 3
- FlightMiles: 3
- FlightTrans: 3
- DaysSinceEnroll: 3

2.1.1.1. Compared to the other clusters, Cluster 1 has the largest average values in which features (if any)? Based on this, how would you describe the Airline's customers in Cluster 1?

Cluster 1:

- Highest average values for Balance, BonusTrans, FlightMiles, FlightTrans, and DaysSinceEnroll.
- Highly active and loyal customers.
- Accumulate a significant number of miles eligible for award travel.
- Engage in numerous non-flight bonus transactions.
- Travel substantial flight distances and make frequent flight transactions.
- Long-standing relationship with the airline.

2.1.1.2. Apply the previous analysis to the remaining four clusters and describe them.

Cluster 2:

- Highest average value for QualMiles.
- Customers with the highest number of miles that qualify for TopFlight status.
- Indicative of high flight activity.
- Significant status level within the frequent flyer program.

Cluster 3:

- Lowest average values for all features.
- These customers are less active, have fewer accumulated miles, and have been enrolled in the frequent flyer program for a relatively shorter duration.

Cluster 4:

- Highest average value for BonusMiles.
- Customers with the highest number of miles earned from non-flight bonus transactions.
- Actively participate in bonus programs.

Cluster 5:

- Highest average value for DaysSinceEnroll.
- Customers enrolled in the frequent flyer program for the longest duration.
- Long-term members who maintain their engagement with the airline.

Selecting the K parameter, The Elbow method

```
[[ 1. 27993. ] [ 2. 21986.77617] [ 3. 18131.37098] [ 4. 15490.86141] [ 5.
13515.25838] [ 6. 12157.9019 ] [ 7. 11044.40422] [ 8. 10110.55239] [ 9.
9472.72598] [ 10. 8926.16985] [ 11. 8450.25224] [ 12. 8053.05877] [ 13.
7619.80681] [ 14. 7374.51932] [ 15. 7096.45098] [ 16. 6846.99845] [ 17.
6660.11762] [ 18. 6487.20065] [ 19. 6314.20032] [ 20. 6130.18058] [ 21.
5992.57697] [ 22. 5848.5045 ] [ 23. 5687.07565] [ 24. 5518.8288 ] [ 25.
5431.58005] [ 26. 5350.70214] [ 27. 5215.27231] [ 28. 5114.74022] [ 29.
5080.68656]]
```

2.1.2.1. Describe the relationship between the parameter k and the obtained sum of squared errors.

The relationship between the number of clusters (k) and the sum of squared errors (SSE) follows a general trend: as k increases, the SSE tends to decrease. The reduction in SSE is significant when going from 1 to 2 clusters, but the rate of reduction slows down as k increases further.

2.1.2.2. Based on the results, indicate the optimal parameter k, that you consider appropriate to clustering the Airlines Cluster dataset.

The plot of the sum of SSE shows a clear elbow point at k = 5, so that must be the optimal value for k.

Hierarchical Clustering

Cluster with highest average value for Balance, QualMiles, BonusMiles, BonusTrans, FlightMiles, FlightTrans, DaysSinceEnroll:

- Balance: 3
- QualMiles: 1
- BonusMiles: 3
- BonusTrans: 2
- FlightMiles: 2
- FlightTrans: 2
- DaysSinceEnroll: 3

Cluster with lowest average value for Balance, QualMiles, BonusMiles, BonusTrans, FlightMiles, FlightTrans, DaysSinceEnroll:

- Balance: 5

- QualMiles: 5
- BonusMiles: 5
- BonusTrans: 5
- FlightMiles: 5
- FlightTrans: 5
- DaysSinceEnroll: 5

2.2.1. Do you expect that Cluster 1 of the Hierarchical clustering to be necessarily similar to KMeans clustering? Compare the results obtained using K-Means and Hierarchical clustering.

No, the two clustering methods employ different approaches to cluster formation, which can lead to distinct results.

We can see that the cluster assignments and average values differ between the two models. The k-means model identified Cluster 1 with the highest average values for several features, while the hierarchical clustering model identified Cluster 3 with the highest average values for similar features. Similarly, the cluster with the lowest average values differs between the two models (Cluster 3 in k-means and Cluster 5 in hierarchical clustering).

2.2.2. Compared to the other hierarchical clusters, Cluster 1 has the largest average values in which features (if any)? Based on this, how would you describe the Airline's customers in Cluster 1? Apply the previous analysis to the remaining four clusters and describe them.

Cluster 1 has the highest average value in QualMiles.

Cluster 1:

- Highest average value for QualMiles.
- Customers with the highest number of miles that qualify for TopFlight status.
- Indicative of high flight activity.
- Significant status level within the frequent flyer program.

Cluster 2:

- Highest average values for BonusTrans, FlightMiles, and FlightTrans.
- They tend to engage more frequently in bonus transactions, accumulating bonus miles at a higher rate.
- These customers also tend to travel more often, accumulating more flight miles and engaging in a higher number of flight transactions.

- They may be frequent flyers or more actively involved in redeeming rewards through flights.

Cluster 3:

- Highest average values for Balance, BonusMiles, and DaysSinceEnroll.
- They are likely long-standing and loyal customers.
- These customers have accumulated a significant balance and earned a substantial number of bonus miles.

Cluster 4:

- Moderate average values across all features. Middle-ground customer segment.
- These customers may represent a diverse group with varying levels of activity and engagement with the airline.

Cluster 5:

- Lowest average values for all features.
- These customers may represent a less active or less engaged segment.

2.2.3. Perform a new hierarchical clustering of the AirlineCluster Dataset using a different linkage method and compare with the clustering obtained using the 'ward' method. Explain what happens.

Cluster with highest average value for Balance, QualMiles, BonusMiles, BonusTrans, FlightMiles, FlightTrans, DaysSinceEnroll:

- Balance: 4
- QualMiles: 1
- BonusMiles: 2
- BonusTrans: 5
- FlightMiles: 5
- FlightTrans: 5
- DaysSinceEnroll: 4

Cluster with lowest average value for Balance, QualMiles, BonusMiles, BonusTrans, FlightMiles, FlightTrans, DaysSinceEnroll:

- Balance: 3
- QualMiles: 4
- BonusMiles: 3

- BonusTrans: 3
- FlightMiles: 3
- FlightTrans: 3
- DaysSinceEnroll: 5

The differences are significant. For example, Cluster 5 has now the greatest amount of high average values while it was the one with the least amount of those when using “ward”. Same happens to Cluster 3 which is also a big difference.

The “average” linking method aims to create clusters with similar average dissimilarities to each other. Lowest average values have been affected because of this.

DBSCAN

2.3.1. The estimated number of clusters and noise points in the three previous scenarios.

***** DBSCAN a) *****

DBSCAN configuration parameters: eps = 0.3, min_samples = 10

Estimated number of clusters: 5

Estimated number of noise points: 2297

***** DBSCAN b) *****

DBSCAN configuration parameters: eps = 0.3, min_samples = 5

Estimated number of clusters: 12

Estimated number of noise points: 1940

***** DBSCAN c) *****


DBSCAN configuration parameters: eps = 0.2, min_samples = 10

Estimated number of clusters: 7

Estimated number of noise points: 3018

2.3.2. Explain how decreasing “eps” affects the estimated number of clusters.

When the value of epsilon in DBSCAN is decreased, the estimated number of clusters tends to increase. Smaller epsilon values create smaller neighborhoods and reduce the connectivity



between points, causing previously connected points to become disconnected and form separate clusters. This increases fragmentation.

2.3.3. Explain how decreasing “min_samples” affects the estimated number of clusters.

Decreasing the value of the min_samples parameter in DBSCAN leads to an increase in the estimated number of clusters. By reducing the minimum number of neighboring points required for a point to be considered a core point, more points are classified as core points, expanding the clusters and potentially merging previously separate clusters. It also decreases the number of noise points because more points are assigned to clusters.