MARKET SEGMENTATION FOR AIRLINES

Market segmentation is a strategy that divides a broad target market of customers into smaller, more similar groups, and then designs a marketing strategy specifically for each group. Clustering is a common technique for market segmentation since it automatically finds similar groups given a data set.

In this problem, we'll see how clustering can be used to find similar groups of customers who belong to an airline's frequent flyer program. The airline is trying to learn more about its customers so that it can target different customer segments with different types of mileage offers.

The file AirlinesCluster.csv (/c4x/MITx/15.071x/asset/AirlinesCluster.csv) contains information on 3,999 members of the frequent flyer program. This data comes from the textbook "Data Mining for Business Intelligence," by Galit Shmueli, Nitin R. Patel, and Peter C. Bruce. For more information, see the website for the book (http://www.dataminingbook.com/).

There are seven different variables in the dataset, described below:

- **Balance** = number of miles eligible for award travel
- **QualMiles** = number of miles qualifying for TopFlight status
- **BonusMiles** = number of miles earned from non-flight bonus transactions in the past 12 months
- **BonusTrans** = number of non-flight bonus transactions in the past 12 months
- **FlightMiles** = number of flight miles in the past 12 months
- **FlightTrans** = number of flight transactions in the past 12 months
- **DaysSinceEnroll** = number of days since enrolled in the frequent flyer program

PROBLEM 1.1 - NORMALIZING THE DATA (2/2 points)

Read the dataset AirlinesCluster.csv (/c4x/MITx/15.071x/asset/AirlinesCluster.csv) into R and call it "airlines".

Looking at the summary of airlines, which two variables have (on average) the smallest values?

- [ ] Balance
- [ ] QualMiles
- [ ] BonusMiles
- [x] BonusTrans ✔
- [ ] FlightMiles
- [x] FlightTrans ✔
- [ ] DaysSinceEnroll

Which two variables have (on average) the largest values?

- [x] Balance ✔
- [ ] QualMiles
- [x] BonusMiles ✔
- [ ] BonusTrans
- [ ] FlightMiles
- [ ] FlightTrans

- [ ] DaysSinceEnroll

EXPLANATION

You can read in data and look at the summary with the following commands:

airlines = read.csv("AirlinesCluster.csv")

summary(airlines)

For the smallest values, BonusTrans and FlightTrans are on the scale of tens, whereas all other variables have values in the thousands.

For the largest values, Balance and BonusMiles have average values in the tens of thousands.

Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.2 - NORMALIZING THE DATA (1/1 point)

In this problem, we will normalize our data before we run the clustering algorithms. Why is it important to normalize the data before clustering?

- ○ If we don't normalize the data, the clustering algorithms will not work (we will get an error in R).
- ○ If we don't normalize the data, it will be hard to interpret the results of the clustering.
- ◉ If we don't normalize the data, the clustering will be dominated by the variables that are on a larger scale. ✔
- ○ If we don't normalize the data, the clustering will be dominated by the variables that are on a smaller scale.

EXPLANATION

If we don't normalize the data, the variables that are on a larger scale will contribute much more to the distance calculation, and thus will dominate the clustering.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 1.3 - NORMALIZING THE DATA (2/2 points)

Let's go ahead and normalize our data. You can normalize the variables in a data frame by using the preProcess function in the "caret" package. You should already have this package installed from Week 4, but if not, go ahead and install it with install.packages("caret"). Then load the package with library(caret).

Now, create a normalized data frame called "airlinesNorm" by running the following commands:

preproc = preProcess(airlines)

airlinesNorm = predict(preproc, airlines)

The first command pre-processes the data, and the second command performs the normalization. If you look at the summary of airlinesNorm, you should see that all of the variables now have mean zero. You can also see that each of the variables has standard deviation 1 by using the sd() function.

In the normalized data, which variable has the largest maximum value?

- ○ Balance
- ○ QualMiles
- ○ BonusMiles
- ○ BonusTrans
- ◉ FlightMiles ✔

- FlightTrans
- DaysSinceEnroll

In the normalized data, which variable has the smallest minimum value?

- Balance
- QualMiles
- BonusMiles
- BonusTrans
- FlightMiles
- FlightTrans
- DaysSinceEnroll ✔

EXPLANATION

After running the two lines of code to normalize the data, you can look at the summary of airlinesNorm with the command:

summary(airlinesNorm)

You can see from the output that FlightMiles now has the largest maximum value, and DaysSinceEnroll now has the smallest minimum value. Note that these were not the variables with the largest and smallest values in the original dataset airlines.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.1 - HIERARCHICAL CLUSTERING (1 point possible)

Compute the distances between data points (using euclidean distance) and then run the Hierarchical clustering algorithm (using method="ward") on the normalized data. It may take a few minutes for the commands to finish since the dataset has a large number of observations for hierarchical clustering.

EXPLANATION

You can compute the distances and run the hierarchical clustering algorithm with the following commands:

distances = dist(airlinesNorm, method="euclidean")

hierClust = hclust(distances, method="ward")

Then, plot the dendrogram of the hierarchical clustering process. Suppose the airline is looking for somewhere between 2 and 10 clusters. According to the dendrogram, which of the following is NOT a good choice for the number of clusters?

- 2 ✘
- 3
- 6 ✔
- 7

EXPLANATION

You can plot the dendrogram with the command:

plot(hierClust)

If you run a horizontal line down the dendrogram, you can see that there is a long time that the line crosses 2 clusters, 3 clusters, or 7 clusters. However, it it hard to see the horizontal line cross 6 clusters. This means that 6 clusters is probably not a good choice.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 2.2 - HIERARCHICAL CLUSTERING (1/1 point)

Suppose that after looking at the dendrogram and discussing with the marketing department, the airline decides to proceed with 5 clusters. Divide the data points into 5 clusters by using the cutree function. How many data points are in Cluster 1?

776

776

Answer: 776

EXPLANATION

You can divide the data points into 5 clusters with the following command:

clusterGroups = cutree(hierClust, k = 5)

If you type table(clusterGroups), you can see that there are 776 data points in the first cluster.

Hide Answer *You have used 1 of 3 submissions*

PROBLEM 2.3 - HIERARCHICAL CLUSTERING (2/2 points)

Now, use tapply to compare the average values in each of the variables for the 5 clusters (the centroids of the clusters). You may want to compute the average values of the unnormalized data so that it is easier to interpret. You can do this for the variable "Balance" with the following command:

tapply(airlines$Balance, clusterGroups, mean)

EXPLANATION

You can compute the average values for all variables in each of the clusters with the following commands:

tapply(airlines$Balance, clusterGroups, mean)

tapply(airlines$QualMiles, clusterGroups, mean)

tapply(airlines$BonusMiles, clusterGroups, mean)

tapply(airlines$BonusTrans, clusterGroups, mean)

tapply(airlines$FlightMiles, clusterGroups, mean)

tapply(airlines$FlightTrans, clusterGroups, mean)

tapply(airlines$DaysSinceEnroll, clusterGroups, mean)

Advanced Explanation:

Instead of using tapply, you could have alternatively used colMeans and subset, as follows:

colMeans(subset(airlines, clusterGroups == 1))

colMeans(subset(airlines, clusterGroups == 2))

colMeans(subset(airlines, clusterGroups == 3))

colMeans(subset(airlines, clusterGroups == 4))

colMeans(subset(airlines, clusterGroups == 5))

This only requires 5 lines of code instead of the 7 above. But an even more compact way of finding the centroids would be to use the function "split" to first split the data into clusters, and then to use the function "lapply" to apply the function "colMeans" to each of the clusters:

lapply(split(airlines, clusterGroups), colMeans)

In just one line, you get the same output as you do by running 7 lines like we do above. To learn more about these functions, type ?split or ?lapply in your R console. Note that if you have a variable named split in your R session, you will need to remove it with rm(split) before you can use the split function.

Compared to the other clusters, Cluster 1 has the largest average values in which variables (if any)?

- [ ] Balance
- [ ] QualMiles
- [ ] BonusMiles
- [ ] BonusTrans
- [ ] FlightMiles
- [ ] FlightTrans
- [x] DaysSinceEnroll ✔

EXPLANATION

The only variable for which Cluster 1 has large values is DaysSinceEnroll.

How would you describe the customers in Cluster 1?

- ( ) Relatively new customers who don't use the airline very often.
- (•) Infrequent but loyal customers. ✔
- ( ) Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ( ) Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- ( ) Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

EXPLANATION

Cluster 1 mostly contains customers with few miles, but who have been with the airline the longest.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.4 - HIERARCHICAL CLUSTERING (2/2 points)

Compared to the other clusters, Cluster 2 has the largest average values in which variables (if any)?

- [ ] Balance
- [x] QualMiles ✔
- [ ] BonusMiles
- [ ] BonusTrans
- [x] FlightMiles ✔
- [x] FlightTrans ✔
- [ ] DaysSinceEnroll

EXPLANATION

Cluster 2 has the largest average values in the variables QualMiles, FlightMiles and FlightTrans. This cluster also has relatively large values in BonusTrans and Balance.

How would you describe the customers in Cluster 2?

- Relatively new customers who don't use the airline very often.
- Infrequent but loyal customers.
- Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions. ✔
- Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

EXPLANATION

Cluster 2 contains customers with a large amount of miles, mostly accumulated through flight transactions.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.5 - HIERARCHICAL CLUSTERING (2/2 points)

Compared to the other clusters, Cluster 3 has the largest average values in which variables (if any)?

- [x] Balance ✔
- [ ] QualMiles
- [x] BonusMiles ✔
- [x] BonusTrans ✔
- [ ] FlightMiles
- [ ] FlightTrans
- [ ] DaysSinceEnroll

EXPLANATION

Cluster 3 has the largest values in Balance, BonusMiles, and BonusTrans. While it also has relatively large values in other variables, these are the three for which it has the largest values.

How would you describe the customers in Cluster 3?

- Relatively new customers who don't use the airline very often.
- Infrequent but loyal customers.
- Customers who have accumulated a large amount of miles, mostly through non-flight transactions. ✔
- Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

EXPLANATION

Cluster 3 mostly contains customers with a lot of miles, and who have earned the miles mostly through bonus transactions.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.6 - HIERARCHICAL CLUSTERING (2/2 points)

Compared to the other clusters, Cluster 4 has the largest average values in which variables (if any)?

- [ ] Balance
- [ ] QualMiles
- [ ] BonusMiles
- [ ] BonusTrans
- [ ] FlightMiles
- [ ] FlightTrans
- [ ] DaysSinceEnroll

EXPLANATION

Cluster 4 does not have the largest values in any of the variables.

How would you describe the customers in Cluster 4?

- ( ) Relatively new customers who don't use the airline very often.
- ( ) Infrequent but loyal customers.
- ( ) Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ( ) Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- (•) Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions. ✔

EXPLANATION

Cluster 4 customers have the smallest value in DaysSinceEnroll, but they are already accumulating a reasonable number of miles.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.7 - HIERARCHICAL CLUSTERING (2/2 points)

Compared to the other clusters, Cluster 5 has the largest average values in which variables (if any)?

- [ ] Balance
- [ ] QualMiles
- [ ] BonusMiles
- [ ] BonusTrans
- [ ] FlightMiles
- [ ] FlightTrans
- [ ] DaysSinceEnroll

EXPLANATION

Cluster 5 does not have the largest values in any of the variables.

How would you describe the customers in Cluster 5?

- (•) Relatively new customers who don't use the airline very often. ✔
- ( ) Infrequent but loyal customers.
- ( ) Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ( ) Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- ( ) Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

EXPLANATION

Cluster 5 customers have lower than average values in all variables.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 3.1 - K-MEANS CLUSTERING (1/1 point)

Now run the k-means clustering algorithm on the normalized data, again creating 5 clusters. Set the seed to 88 right before running the clustering algorithm, and set the argument iter.max to 1000.

How many clusters have more than 1,000 observations?

2

2

Answer: 2

EXPLANATION

You can run the k-means clustering algorithm with the following commands:

set.seed(88)

kmeansClust = kmeans(airlinesNorm, centers=5, iter.max=1000)

And you can look at the number of observations in each cluster with the following command:

table(kmeansClust$cluster)

There are two clusters with more than 1000 observations.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 3.2 - K-MEANS CLUSTERING (1/1 point)

Now, compare the cluster centroids to each other either by dividing the data points into groups and then using tapply, or by looking at the output of kmeansClust$centers, where "kmeansClust" is the name of the output of the kmeans function. (Note that the output of kmeansClust$centers will be for the normalized data. If you want to look at the average values for the unnormalized data, you need to use tapply like we did for hierarchical clustering.)

Do you expect Cluster 1 of the K-Means clustering output to necessarily be similar to Cluster 1 of the Hierarchical clustering output?

- Yes, because the clusters are displayed in order of size, so the largest cluster will always be first.
- Yes, because the clusters are displayed according to the properties of the centroid, so the cluster order will be similar.
- (selected) No, because cluster ordering is not meaningful in either k-means clustering or hierarchical clustering. ✔
- No, because the clusters produced by the k-means algorithm will never be similar to the clusters produced by the Hierarchical algorithm.

EXPLANATION

The clusters are not displayed in a meaningful order, so while there may be a cluster produced by the k-means algorithm that is similar to Cluster 1 produced by the Hierarchical method, it will not necessarily be shown first.

Hide Answer *You have used 1 of 1 submissions*

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion

New Post

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.