Marketing Analytics Tutorial 3

Market Analytics

Part A) Market Basket Analysis

Exercise 1

We will work with a real data set[1] containing 88.162 customer transactions from a Belgian supermarket chain. Each row of the data set contains an individual's market basket of items that were purchased together. Thereby, items are assigned an unique but arbitrary number, starting with the items observed in the first basket and consecutively adding numbers as required for the following transactions.

First of all, load the data into R from the web using the following command.

```
retail_raw <- readLines("http://fimi.ua.ac.be/data/retail.dat")
```

a) First, convert the raw data into a list, where each element contains the items of a single market basket. Then, the resulting list can be transformed to a formal transactions object, as required for the analysis of association rules.
How many unique items does the supermarket offer? What is the most popular item? What are the sizes of the smallest, largest, and median market basket?
Hint: Use `strsplit()` *to create the list and* `as()` *to obtain the transaction object*.

b) Use `apriori()` to find association rules in the transaction data. Consider tuning the parameters so that the results only consider sets found in at least 0.1% of transactions. Also, confidence should be at least 0.4.
Plot rule confidence against support and interpret the graph.

c) Extract the 50 rules with the highest lift, inspect and plot them separately.

d) To determine the profitability of transactions, simulate margins for each of the items included in the transaction data. Do so by drawing randomly from the N(0.3, 0.3) distribution and storing the results in a separate data frame.
What is the supermarket's margin for a purchase of the items 696 and 699? What is the margin achieved through the 100th transaction in the data set?

Exercise 2

We will work with a simulated data set[2] containing the consumer segment assignments for 300 respondents to a survey of a subscription-based service. The variables are described in the following table:

Variable	Description
`age`	In years

[1] This data set was donated by Tom Brijs and can be found here http://fimi.uantwerpen.be/data/. Please refer to Brijs, T. et al. (1999) "The Use of Association Rules for Product Assortment Decisions: A Case Study" in Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, p. 254 – 260, for a description of the data set.

[2] This data set was taken from "R for Marketing Research & Analytics" by Chris Chapman & Elea McDonnell Feit, http://r-marketing.r-forge.r-project.org/data/.

gender	Binary gender (male or female)
income	Annual income in USD
kids	Number of children
ownHome	Factor variable whether the respondent owns his/her current home
subscribe	Factor variable whether the respondent is a subscriber to the offered service
Segment	Factor variable for consumer segment type (Moving up, Suburb mix, Travelers, Urban hip)

First of all, load the data into R from the web using the following command.
`segdata <- read.csv("http://goo.gl/qw303p", stringsAsFactors = TRUE)`

a) Convert the numeric variables in the data set to ordered factor types instead. More specifically, age should have 5 levels: 19 – 24, 25 – 34, 35 – 54, 55 – 64, and 65+, income 3: Low (<40.000), Medium (40.000 – 70.000), and High (70.000+), and kids should have 4 levels: None, 1, 2, and 3+.
Hint: Use cut( ).
b) Use apriori( ) to find association rules in the segment data. Consider tuning the parameters so that the results only consider sets found in at least 10% of transactions. Also, confidence should be at least 0.4.
Plot rule confidence against support and interpret the graph.
c) Extract the 35 rules with the highest lift, inspect and plot them separately.
d) Find only those association rules with the Urban hip segment on the right side. Also, lift should be at least 1. Sort the rules according to lift.

Part B) Segmentation

We will use a fictional data set[3] containing customer information for some music subscription service for all exercises of this part. The variables are described in the following table:

Variable	Description
age	In years
sex	Binary gender (male or female)
householdIncome	Household annual income in USD
milesDrive	Annual total miles driven by car
kidsAtHome	Number of children (<18 years) living in the household
commuteCar	Dummy variabe whether the respondent regularly commutes by car
drivingEnthuse	Reported enthusiasm for driving (1 – 7)
musicEnthuse	Reported enthusiasm for music (1 – 7)
subscribeToMusic	Factor variable whether the respondent is a subscriber to the service
Segment	Factor variable for customer segment type (CommuteNews, KidsAndTalk, LongDistance, MusicDriver, NonCar, Quiet)

[3] This data set was taken from "R for Marketing Research & Analytics" by Chris Chapman & Elea McDonnell Feit, http://r-marketing.r-forge.r-project.org/data/.

First of all, load the data into R from the web using the following command.
`musicdata_raw <- read.csv("https://goo.gl/s1KEiF", stringsAsFactors = TRUE)`

Exercise 3
Use the music subscription data set to explore distance-based clustering approaches.

a) Inspect the data set and create a copy that does not contain the respondent's segment information. Use this copy for the following tasks.
b) Apply hierarchical clustering (`hclust`) to find groups in the data set and visualize the result. Based on this, what do you think is an appropriate number of segments? Do you find this result useful?
c) Instead, use mean-based clustering (`kmeans`) to find four groups in the data. Do you find these results useful? To illustrate, plot two continuous variables by segment.
d) Plot the clusters found in c) by their principal components with `clusplot( )` and interpret the graph.

Exercise 4
Use the music subscription data set to explore model-based clustering approaches.

a) Apply model-based clustering (`mclust`) to find groups in the data set and visualize the result. What is the suggested number of clusters and are they well-differentiated? Compare to the mean-based solution from 3c).
b) Fit a model-based solution by pre-specifying the number of clusters (G = 2 and G = 4). Compare both results to the suggestion from a).
c) In preparation for latent class analysis, recode all variables to binary factors. Exclude `milesDrive` and `drivingEnthuse`. Split the remaining variables at the specified cut-off values: `age` (<30), `householdIncome` (<55.000), `kidsAtHome` (>0), and `musicEnthuse` (>4).
d) Fit categorical LCA (`poLCA`) with both 3- and 4-class solutions to the data set and visualize the results. Discuss the differences in respondent assignment based on the two solutions and compare their usefulness.

Exercise 5
Use the music subscription data set to explore classification methods.

a) Split the initial data set that includes segment assignment into a training (65%) and test (35%) set.
b) Fit a Naïve Bayes model to predict segment membership based on all of the included variables in the training data set. Assess model performance on the test data comparing its performance to chance.
c) Instead, fit a random forest model to predict segment membership. What is the out-of-bag error rate? Assess model performance on the test data comparing its performance to chance.
d) What variables are most important for the prediction made in c)?
e) Now, fit a random forest model to predict subscription status. How well does it predict the test data and which variables are most important?