Assignment on Clustering & Association Rules

Machine Learning
Prof. Dr. Marc Pouly & Dr. Tim vor der Brück
Lucerne University of Applied Sciences and Arts

Deadline: March 26, 2018.

Exercise 1: Theory Questions (1 Point)

Fill in the online theory quiz on ILIAS.

Acceptance criterion: This is an individual exercise. You can participate in the quiz as many times as you want. ILIAS registers your best result over all runs. You passed this exercise if your best score reaches at least 80%.

Exercise 2: Clustering (1 Point)

This is a team exercise; hand in your solution as a Jupyter notebook by e-mail to Tim vor der Brück no later than the above deadline.

STUcard provided us with an anonymized dump of their member database. It contains >65K real member profiles randomly sampled from a total of >300K member profiles. We are given access to the following information:

- Geschlecht: 0 = M, 1 = W
- Akademiker: 0 = obligatorische Schulbildung, 1 = Matura, 2 = FH, 3 = Uni / ETH
- Auto: 0 = nein, 1 = ja
- Wohnsituation: 0 = Hotel Mama, 1 = WG, 2 = eigene Wohnung
- Online Affinität: Verhalten auf STUcard.ch Website & Mobile App (Tracking)

The STUcard.ch marketers plan to develop marketing campaigns for 5 different target groups, while the number five is primarily motivated by the available marketing budget. Our task is to identify these target groups and to support specification of the different campaigns by investigating the characteristics of each target group. Proceed as follows:

1. Report on data quality
2. Generate 5 clusters; do not forget to normalize (min-max) !
3. Display cluster centers together with the size of each cluster. Hint: it is very hard to interpret normalized results. Therefore, you may want to undo normalization of the cluster centers prior to discussion. Also, you may want to change the scale of the online affinity attribute, e.g. express in percentage rather than absolute value.
4. Draw detailed conclusions about the results; can you derive recommendations for the specific campaign of each target group?
5. Finally, you may want to check whether the assumption of 5 clusters is reasonable. Implement the elbow method and seek for the optimal number of clusters in this data.

Note: You may want to conduct a similar analysis for the Schifffahrtsgesellschaft Vierwaldstättersee. Likewise, our friends from AutoScout24 could use clustering to investigate the different classes of cars that are being traded over their platform. With hundreds of new ads every day, it is clearly impossible to assess such information manually.

Exercise 3: Association Rules (1 Point)

This is a team exercise; hand in your solution as a Jupyter notebook by e-mail to Tim vor der Brück no later than the above deadline.

3.1 Market Basket Analysis

You are given transaction data that we borrowed from Kaggle.

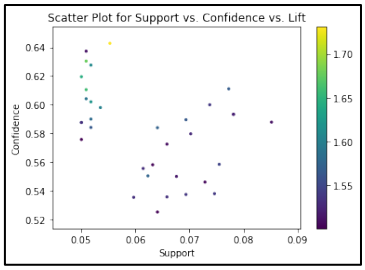
1. Data Wrangling: Create a table of transactions with date and number of items as depicted in the following image

1	[pork, sandwich bags, lunch meat, all- purpose...	2000-01-01	19
2	[toilet paper, shampoo, hand soap, waffles, ve...	2000-01-01	23
3	[soda, pork, soap, ice cream, toilet paper, di...	2000-01-02	31
4	[cereals, juice, lunch meat, soda, toilet pape...	2000-01-02	6
5	[sandwich loaves, pasta, tortillas, mixes, han...	2000-01-02	27
6	[laundry detergent, toilet paper, eggs, toilet...	2000-01-02	28
7	[individual meals, paper towels, tortillas, ve...	2000-01-03	24
8	[ice cream, juice, paper towels, waffles, soda...	2000-01-04	9
9	[juice, poultry, coffee/tea, coffee/tea, dishw...	2000-01-04	5
10	[ketchup, coffee/tea, toilet paper, pork, flou...	2000-01-05	34

2. Generate a histogram on the number of items per transaction; generate a time line on the number of transactions per day. Do people buy more on Saturdays?

3. Use Apriori to generate association rules; identify 3 rules that you consider interesting based on their values for support, confidence and lift. Interpret the values of support, confidence and lift for these specific rules.

4. Display the rule set graphically as below



3.2 Analysis of User Contests

Have you ever heard of Tinder? ☺ Well, our case is not about online dating but conceptually similar. The folks from STUcard.ch launched an online contest for market research. Participants were asked to like (stored as 1) or dislike (stored as 0) 18 images in random order. Unfortunately, we do not have access to these images – still we can come to quite interesting conclusions using the very same techniques as in market basket analysis. Can you find highly significant rules of the kind «*People who liked image X and Y and … also liked image Z*». Again, select the three best rules you can find and interpret them in terms of support, confidence and lift.