Exercise Questions, **Artificial Intelligence**, Part
Knowledge/Discovery/Data Mining

Christoph Schommer

September 10, 2019

1 Association Discovery

1. The *Freie University Berlin* would like to open a supermarket, exclusively for students. To do so, the President asks you how to do on the basis of the following First-Day-Customer-Data (in German).

```
TID    Items
===    ===========
#100    Cola, Taschentuch, Bier, Bananen, Wurst A, Gouda, Sellerie, Zitronen
#110    Puder, Bananen, Toilettenpapier, Gouda, Pizza, Schololade, Eiscreme
#200    Bier, Toilettenpapier, Babycreme, Puder, Pizza, Bananen
#410    Eiscreme, Puder, Salami, Steak Argentina, Scotch, Seife, Zahnpasta
#440    Bier, Puder, Pizza, Toilettenpapier, Bananen, Duschgel ‘‘Monster’’
#650    Pizza, TicTacs, Puder, Maoam, Haribo, Axe ‘‘MakeMeATiger’’, Wein
#700    Puder, Pizza, Gouda, Bananen, Toilettenpapier
#800    Pizza, Bananen, Toilettenpapier, Puder
#900    Spueli, Puder, Gouda, Bier, Pizza
#950    Pizza, Bier, Toilettenpapier, Puder, Bananen
```

   Define (by yourself) an order, a minimum confidence and a minimum support and apply the apriori-algorithm to underline your decision. Explain how you organise (product placement) the new supermarket.

2. Which of the following statements is correct and why?

   (a) if x→ y as well as y→ z are valid association rules, then also x→ z (transitivity).

   (b) With apriori, it is possible to create the following rule: x→ x (reflexivity).

   (c) $\forall x, y$: $0 \leq \text{Lift}(x \rightarrow y) \leq 1$

   (d) The confidence parameter is commutative: Confidence(x→y) = Confidence(y→x)

   (e) The lift parameter is commutative: Lift(x→y) = Lift(y→x)

2 Clustering

1. Given the following data situation (taken from a weblog file):

   Carole has visited the following web pages in that order:
   banana.html → baby.html → banana.html → help.html → fashion.html → kids.html

   Julie has visited the following web pages in that order:
   baby.html → help.html → banana.html → banana.html → fashion.html → kids.html

   Friedrich has visited the following web pages in that order: help.html → banana.html → banana.html → kids.html

   Maximilian has visited the following web pages in that order: sports.html → baby.html → kids.html → fashion.html → help.html

   where fashion.html contains information about the newest clothes, baby.html and kids.html some information about baby and kids products, respectively, banana.html offerings concerning the best bananas in the world, sports.html some information about the newest sports collection, and help.html some general information about the shop.

   Your tasks (write down all intermediate steps and explain your answers!):

   - Introduce the following attributes: **Name-of-Visitor**, **Number-of-Different-Pages**, **Start-Page**, **End-Page**, **Visited-a-Fruit-Page?**, **Visited-a-webpage-more-than-once?**
   - For **Number-of-Different-Pages**: discretise the original values either to low (if < 3) or to high (if ≥ 3)
   - Then, state the database table with all new attributes and the database records.
   - Finally, create a distance (and similarity) matrix by applying the **Hamming Distance**. Explain, who of the users may belong to the same clusters.

2. Given the following data points: x_1=(0,0,0); x_2=(1,0,0); x_3=(1,0,1); x_4=(1,1,0); x_5=(1,1,1); x_6=(0,0,1); x_7=(0,1,0); and x_8=(0,1,1). Explain, which data points belong to which cluster under the assumption that you use a k-means algorithm with k = 2 and the randomly selected centroids c_0=(0.5,0.5,0) and c_1=(0,0.5,0.5), respectively.