Data and Web Science Group
B6 – B1.15
68159 Mannheim

Data Mining

Exercise 8: Association Analysis

8.1. Analyzing the Shopping Basket Data Set

1. The Shopping Basket data set is provided as an Excel file in ILIAS. Load the data set with the *read_excel* function.

2. Mine frequent item sets from the data set using the *frequent_itemsets* function (support = 0.2). Which items are usually bought together with the laptop, the netbook and the printer?

3. Create association rules from the frequent item sets using the *association_rules* function with a confidence of 0.7. What do the rules tell you about the relationship between Asus EeePC netbooks, 2 GB DDR3 RAM extensions and Netbook Schutzhüllen? What do the lift values (compute it with the *rules_stats* function) tell you about the interestingness of the rules?

8.2. Finding Frequent Pattern in the Adult Data Set

1. Import the Adult-tweaked data set and print a description of the dataset using the *describe* function. The Adult-tweaked data set is provided on ILIAS as an ARFF file.

2. Prepare the data set for Frequent Pattern Mining by: 1) reducing the size of the data set to 5000 examples using sampling; 2) using a subset of attributes: age, workclass, education, occupation, race, sex, hours-per-week, native-country and class; 3) discretizing the attributes age and hours-per-week into three user defined ranges (think about ranges that could make sense for the attributes). How many attributes does the resulting data set have?

3. Find frequent item sets that have a support above 0.2. What can you learn from these item sets about the people who earn less than 50K a year?

4. Given the large number of examples and the low min-support threshold, the number of frequent item sets containing the *education* attribute is surprisingly low. Moreover, only one value for this attribute is present in the resulting frequent item sets. Why is this the case? How could you aggregate the data to change this without losing too much information? Also look at the *native-country* attribute and think of a possible aggregation for it.

5. Restrict the frequent item sets to the ones containing “class = >50K” and lower the support so that a decent number of item sets is discovered. What can you learn from these item sets about the people who earn more than 50K a year?

8.3. Mining Association Rules from the Adult Data Set

1. Create association rules from the frequent item sets from exercise 8.2.3 using the *association_rules* function. Which rules do you consider interesting? Consider both =>50K and <50K classes.

2. Now we want to focus on the relationship of the occupation, education and income of immigrants: instead of sampling the data set, filter the data set: native-country != United-States. Which rules do you consider interesting?

ChatGPT Bonus Exercises

Reminder: Do not take the answers of ChatGPT at face value! Always cross-check with lecture slides, literature and/or the teaching staff!

C.1. Discuss Measures and Quality of Rules

- Discuss Association Analysis (Association Rule Mining) with ChatGPT. How are the metrics support, confidence, and lift used to evaluate the quality of discovered association rules? Discuss the following three scenarios and identify the scenario that likely indicates an interesting rule:

  1. Low support and confidence as well as a lift close to 1.
  2. High support and low confidence and a lift close to 1.
  3. High support and confidence as well as a lift greater than 1.

C.2. Generate code to compare the Apriori algorithm and the FPgrowth algorithm

- In the lecture you learned about the Apriori algorithm for generating frequent itemsets. In the exercise we have used the FPgrowth algorithm instead to produce such itemsets. Have ChatGPT explain the difference between both algorithms to you. Ask ChatGPT for some code to produce frequent itemsets using the apriori algorithm with the orange3 library.

- Will both algorithms produce the same itemsets? Discuss this question with ChatGPT and ask for code that applies both algorithms using the orange3 library and subsequently compares the generated frequent item sets. Apply the code for one of the datasets from the exercise.

C.3. Self-Assessment

- Ask ChatGPT to create a pen and paper exercise by giving you some transaction items to practice to calculate the confidence of some association rules.

- Ask ChatGPT to generate some multiple-choice questions for graduate students related to association analysis. Include questions regarding pre-processing nominal and continuous features and shortcomings of the metrics support, confidence and lift that are used to assess the relevancy of association rules.