Rayanna Harduarsingh

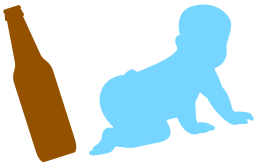
October 21st, 2020

IST 687

Module 9 Notes

Chapter 17/Module 9- Data Mining

- **Dating Mining**
  - The use of algorithms and computers to discover novel and interesting patterns within data
  - Example: Beer and Diapers



    - Supermarket that analyzed patterns of purchasing behavior and found that diapers and beer were often purchased together. The supermarket manager decided to put a beer display close to the diaper aisle and supposedly sold more of both products as a result.
  - Includes four processes:
  - (1) Data Preparation (time consuming)
    - Making sure data is organized in the right way and transforming it
    - Missing data fields are filled in
    - Inaccurate data are located and repaired or deleted
    - Data is recoded as necessary to make them amenable to the kind of analysis we have in mind
  - (2) Exploratory Data Analysis
    - Getting to know the data using histograms and other visualization tools
    - Looking for preliminary hints that will guide our model choice
    - Exploration process also involves figuring out the right values for key parameters

    - (3) Model Development
        - Choosing and developing a model (most complex and most interesting of the activities of a data miner)
        - Here is where where you test out a selection of the most appropriate data mining techniques.
        - Depending on the structure of a data set, there could be dozens of options, and choosing the most promising one has as much art in it as science.
    - (4) Interpretation of Results
        - Focuses on making sense out of what the data mining algorithm had produced
        - This is the **most important step** from the perspective of the data user, because this is where an actionable conclusion is formed.
            - In the example of beer and diapers, the interpretation of the association rules that were derived from the grocery purchasing data is what led to the discover of the beer–diapers rule and the use of that rule in reconfiguring the displays in the store.
- Question: How is this different from the overall data science process?
    - I think in this process, it incorporates everything we have learned so far such as compressing and organizing the data into something readable, visualizing the data, and drawing some educated conclusions. It's a combination of everything and every skill in data science where you bring all those skills together and use it in data mining that can aid in consumer insights, trends, and improve marketing tactics.

- **Associative Rule Mining**
    - Association rules are if/then statements.
    - Help uncover relationships between seemingly unrelated data.
        - Ex: If a customers buys a dozen eggs, that person is 80% likely to also purchase milk
    - RStudio
        - Need to download "arules" package

- Support rule
    - Proportion that a paring occurs across all baskets
        - # of rows having both A AND B / Total # of rows
- Confidence quantity
    - How frequently a particular pair occurs among all the items when the first item is present
        - # of rows having both A AND B / # of rows with A
- Lift: Confidence / Probability of Second Item
    - Confidence / Expected Confidence
- Question: Real World Examples / Possible Issues
    - Examples would include everyday shopping of course. For example, when buying a new iPhone, a user will be more likely to buy a phone case or a screen protector. Another example would be predicting trends, so for example, in the winter, people will be looking to buy winter jackets and sweaters. Another example would be taken from streaming services. If a user is into comedies and have been watching a lot of comedy movies, they can be recommended more funny movies. However, some issues than can occur is that some associative rules may not be entirely accurate or directly correlated. Some other factors can be incorporated that we may not know of or cannot see that results in the final outcome.

RStudio for Associative Rule Mining:

- > data(Groceries)
    - Making a dataset ready
- > itemFrequencyPlot(Groceries,support=0.1)
    - Produces a plot
    - Based on summary outcome in textbook, the item “yogurt” appeared in 1,372 out of 9,835 rows, or in about 14% of cases. So we can set the support parameter to somewhere around 10% and 15% in order to get a manageable number of items

- > itemFrequencyPlot(Groceries, support=0.05,cex.names=0.5)
    - cex.names reduces the font size on the labels
- ruleset <- apriori(Groceries, parameter = + list(support = 0.01,confidence = 0.5))
    - Apriori uses an iterative approach known as level-wise search

Can you explain more about the use of Apriori and its purpose?