**D212 Task 3: Association Rules and Lift Analysis**

**Joseph Duszynski**

**ID #011237236**

**A.1. Proposal of question**

Using the provided medical market basket dataset (WGU, 2024 [1]), this project seeks to answer the following question: “Can a market basket analysis of patient prescriptions, using the Apriori algorithm, find meaningful associations of prescriptions commonly used together? Specifically, which medications are most frequently used in combination with Diazepam?”

**A.2. Defined goal**

This project aims to find worthwhile and insightful associations between patient prescriptions frequently used together. In particular, it looks to find medications most commonly used with Diazepam such that their conditional frequencies are increased from their natural frequencies (throughout the dataset) by a lift factor of sufficient magnitude. These findings can assist hospital medical staff in reviewing a predetermined list to diagnose a patient’s need for other medications should they be found to need Diazepam (as well as other medications). Hospital pharmacists can also benefit by having their medications efficiently organized much like a grocery store where commonly bundled items are found near one another.

**B.1. Explanation of market basket**

Market basket analysis seeks to determine sets of items that are commonly purchased (or used) together with respective conditional probabilities, forming what are referred to as “association rules”. For example, customers who purchase item X might purchase item Y with a probability of 0.6, which would go unnoticed through an analysis of inventory sold by item name. This understanding informs organizations how to rearrange item positions and product recommendations to increase revenue and/or improve customer experience.

Market basket analysis takes customer “baskets” of products as observations (e.g. customer 1 purchased items #2, #6, and #8; customer 2 purchased items #3, #6, and #12; and so forth) and transforms their purchases into a dataset with the distinct products as features and Boolean values recording if a given customer (observation) purchased that product.

This transformed dataset is used as input for the Apriori algorithm, which takes a minimum support (frequency) and returns sets of items (some of which may only consist of a single item) with their respective frequencies if they’re above that threshold.

These frequencies are then used to calculate conditional probabilities and related statistics for antecedent-consequent pairs in a Bayesian-influenced way. The confidence for such a pair is defined as the frequency (support) of that pair divided by the antecedent’s frequency, reflecting the likelihood a customer would purchase the consequent item(s) given they purchased the antecedent. The lift is the ratio of the confidence to the frequency of the consequent, the multiplicative factor by which a customer would purchase the consequent compared to the overall purchasing rate of the consequent item(s) with no conditional assumptions. The leverage is the difference between the frequency of the antecedent-consequent pair and the product of the antecedent’s frequency and consequent’s frequency, a measure of the discrepancy in the observed frequency of that pair to one where the purchases are independent with no conditional probabilities. The conviction is defined as .

The resulting outcomes from the analysis provides the user with a table of antecedent-consequent pairs as described above (along with supports, confidence, lift, leverage, and conviction). The user can then select the most relevant association rules on the basis of support (for each antecedent-consequent pair), confidence, lift, leverage, conviction, another metric of their choosing, or a combination of thresholds and conditions, that suits their project needs.

**B.2. Transaction example**

One example of a transaction in this dataset is: amlodipine, diazepam, clopidogrel, metoprolol succinate XL, triamterene, fenofibrate, cialis. Below is how it’s presented from the dataframe:

print(df\_med.iloc[16])

Presc01 amlodipine

Presc02 diazepam

Presc03 clopidogrel

Presc04 metoprolol succinate XL

Presc05 triamterene

Presc06 fenofibrate

Presc07 cialis

Presc08 NaN

Presc09 NaN

Presc10 NaN

Presc11 NaN

Presc12 NaN

Presc13 NaN

Presc14 NaN

Presc15 NaN

Presc16 NaN

Presc17 NaN

Presc18 NaN

Presc19 NaN

Presc20 NaN

Name: 33, dtype: object

The NaN values imply this patient only purchased seven prescriptions (with a ceiling of 20 for any given transaction).

**B.3. Market basket assumption**

Market basket analysis determines frequently occurring groups of items with the assumption that the purchases for that transaction have potential dependent relationships between one another to be uncovered through the collection of many transactions. That is, customer purchases aren’t random.

Additionally, it’s time-agnostic (without modifications) nor can it distinguish a set of 1000 purchases by 1000 customers from 100 customers each making 10 purchases. One individual who routinely purchases ice cream and sea bass could distort any conclusions if they were extrapolated to other grocery stores. Depending on the timeframe in which the data is collected, there may be a seasonal relationship that is irrelevant at other times. For example, marshmallows, graham crackers, chocolate, and lighter fluid may be associated during the summer for bonfires and s’mores, whereas purchases of marshmallows aren’t likely to have a strong association to graham crackers in the winter in northern climates. Although it depends on how the data is collected, one customer making separate trips or transactions due to forgetting something, having an issue with completing the purchase, or some other reason, will not have any detected associations as they might as well be two separate people.

**C.1. Transforming the data set**

See the attached file “medical\_cleaned\_market\_basket.csv” for the cleaned and transformed data as produced by TransactionEncoder().

The initial dataset has null rows in every other row, so these are removed with:

# Every other row is null, so those are removed here  
df\_med = df\_med.dropna(axis=0, how='all')

The data is then prepared using nested\_lists() to create a 2D array of patients’ prescriptions (the first axis is the patient number and the second is a list of their prescriptions with length between 1 and 20 (nulls are omitted)). This 2D array is passed to transactions\_and\_rules() with specified minimum thresholds and a selected metric (used later in the function) to create a dataframe of Boolean values with the 119 unique medications as columns and 7501 patients as observations. This resulting dataframe is the transformed dataset produced by TransactionEncoder() mentioned above.

See below:

A screenshot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

**C.2. Code execution**

See the attached “market\_basket.py”. Using the following:

A screenshot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

It returned the following output:

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

The output from commands reducing the association rules table by condition will be show in sections C.4. and D.1.

**C.3. Association rules table**

The association rules table provided in the previous section lists support, lift, and confidence values for antecedent-consequent pairs:

A screenshot of a computer screen

Description automatically generated

**C.4. Top three rules**

Finding the most relevant association rules with and , sorted in descending order by lift:

A screenshot of a computer screen

Description automatically generated

Keeping in mind the combination of lisinopril, carvedilol, and abilify may be considered a duplicate for certain purposes, the top three (with one potentially redundant) rules are:

A screenshot of a computer screen

Description automatically generated

A moderate confidence of 0.25 is a reasonable cutoff to ensure patients using the consequent’s medications can be addressed while safeguarding against conclusions and actions for a weak association. A threshold of 1.8 for the lift filters the rules to those with a notable increase from the independent probability of the consequent to its conditional (on the condition the antecedent is used). A minimum support threshold of 0.015 was used as a compromise to capture relevant patterns while avoiding noise as much as possible.

From this dataset, the most important rule on the basis of lift (satisfying the above thresholds) is methylprednisone and lisinopril with a lift of 3.29 and a confidence of 0.32.

The second most important rule is split between (carvedilol, abilify) and lisinopril or its alternate (lisinopril, abilify) and carvedilol. The first has lift 2.91 and confidence 0.29 while the second has lift 2.39 and confidence 0.42.

The third most important rule is metformim and atorvastatin, with lift 2.32 and confidence 0.30.

**D.1. Significance of support, lift, and confidence summary**

Continuing with the output of

df\_diaz\_red, df\_diaz = rules\_by\_med('diazepam', tr\_rules, 'lift', 'confidence')

As the primary interest is in bundles of medications with significant lift values, duplicates where Diazepam occurs in the antecedent or consequent are not of primary interest, so the ordering with the highest confidence was selected.

A screenshot of a computer screen

Description automatically generated

As explained in section B.1., the support for an antecedent-consequent pair represents its overall frequency across the dataset. With 119 medications used in this dataset and over 50% of patients having no more than three medications, the support values are expected to be quite low overall. However, extraordinarily low values could occur due to just a handful of patients using certain combinations of medicines, so a balanced minimum threshold of 1.5% (at least ~110 patients) was used in this project.

The confidence is the conditional probability a patient will also use the consequent for a given antecedent. For purposes of assisting patients and pharmacy organization, a threshold value of 0.25 is adequate without being too restrictive nor too lenient.

The lift is the multiplicative difference between the confidence and the support (frequency) of the consequent. A value of 1.0 implies the consequent’s rate of use is no different from the general population of the dataset. Restricting this to 1.8 retains association rules with a notable increase (multiplicatively) in conditional probability (compared to the baseline throughout the dataset).

The support, lift, and confidence must all be monitored before forming any conclusions, as outliers, noise, and thresholds that are either too strict or too lenient can provide meaningless results. See section C.4. for specific values of the top three rules.

**D.2. Practical significance of findings**

Diazepam is prescribed at a rate of 16.38% across the patients in this data. Given its prevalence, knowing which medications are commonly associated with it can assist hospital pharmacists in efficient organization while also alerting physicians and patients to other medications they may need or wish to consider. In some cases, the associations could prompt diagnostic tests and examinations that uncover previously unknown health issues.

Reviewing the table for Diazepam in section D.1., the top four results all have lifts of at least 1.5 and confidence values of 0.24. Notably, atorvastatin and diazepam have a support of 3.21%, significantly higher than most other antecedent-consequent pairs. The confidence and lift values are moderately increased. Realistically, these results will be of greatest relevance to pharmacists organizing their inventory. Diazepam carries a higher confidence when it’s the consequent in most cases, implying it tends to be a “secondary” medication applied with others. That is, physicians likely won’t need to further diagnose patients for other conditions as that process was likely already done sufficiently well.

For the top three rules overall from section C.4., the antecedent supports are moderately high, the confidence values are sufficient, and the lift values are considerably high. Enough patients are affected that physicians and pharmacists may benefit from this information, although in all likelihood they’re already aware of the associations.

**D.3. Course of action**

The goal of this project was finding frequently associated combinations of medications across the dataset as well as those associated with Diazepam in particular. Given what was discussed in section D.2. for the practical relevance of the results, there’s a strong chance that physicians and pharmacists are already aware of these associations and have no need to make any adjustments. However, at the risk of “data confirms the obvious”, the organization could release a brief summary of the findings to relevant staff. Pharmacy stock would be better organized placing Diazepam near Abilify, Carvedilol, Amlodipine, Glipizide, and others. Lisinopril is commonly used with methylprednisone, Abilify, and Carvedilol. One potentially unknown relationship to physicians is 40.5% of patients taking Diazepam (treats muscle spasms and anxiety) and Carvedilol (a beta blocker) also take Abilify (an antipsychotic). Diazepam lists paranoia and other mental health problems as side effects, but the high percentage of patients requiring an antipsychotic while taking these two medications isn’t necessarily immediately obvious or widely known.

**E. Panopto video of code**

See the attached link: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3523afdc-1d01-4de3-999d-b17d00dec2a4

**F. Sources for third-party code**

**1.** WGU. 2024. D212 Data Mining II “Data Sets and Associated Data Dictionaries”. Medical Data and Dictionary Files. Retrieved May 23, 2024, from <https://access.wgu.edu/ASP3/aap/content/jf8rcds032ldktfces9r.html>.

**2.** Kamara, Kesselly. 2024. WGU D212 “Data Mining II – D212 Task 3”. Retrieved May 26, 2024, from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=db85c4f1-0da5-4bde-a1a4-b07c0019d46d.

**G. Sources**

No other sources were used.