Pattern Mining

Dr. Matthieu Cisel

February 2023

1 Goals

In this class, students will explore a wide diversity of pattern mining techniques .

1. Correlation Mining
2. Market Basket Analysis
3. String and Itemsets Sequential Mining
4. Episode mining
5. Periodic pattern mining
6. Spatial associations
7. Text mining (quality phrases)

2 Instructions

2.1 Datacamp and Coursera classes

You must have completed by the end of the course both Market Basket Analysis with Python in Datacamp, and the weeks 1, 2 and 3 of Coursera's Pattern Mining class (University of Illinois). For the n-grams section, you can follow videos on Datacamp in the Bag-of-Words classes (R or Python).

2.2 Expectations

At the end of the learning unit, we expect you to submit :

1. A first PDF and an html version of your notebooks (one notebook per section). A zip encompassing all of your figures and notebooks is required. Your name must appear in the name of the document. Code annotation is mandatory.

2. A short report including the description of 1/2 figures that you have made (the choice of the figure is yours), with a proper interpretation on one hand, and a fallacious interpretation on the other hand. The choice of the fallacy is yours, but you must explain why it is a fallacy.

3. A certificate of the Market Basket Analysis class in Python (Datacamp)

3 Correlation mining

The NBA dataset provides a wide array of statistics on key players. The signification of the corresponding variables is given below.

GP – Games Played
W – Wins
L – Losses
MIN – Minutes Played
FGM – Field Goals Made
FGA – Field Goals Attempted
FG% – Field Goal Percentage
3PM – 3 Point Field Goals Made
3PA – 3 Point Field Goals Attempted
3P% – 3 Point Field Goals Percentage
FTM – Free Throws Made
FTA – Free Throws Attempted
FT% – Free Throw Percentage
OREB – Offensive Rebounds

Figure 1: Glossary for the NBA dataset

DREB – Defensive Rebounds
REB – Rebounds
AST – Assists
TOV – Turnovers
STL – Steals
BLK – Blocks
PF – Personal Fouls
FP – Fantasy Points
DD2 – Double doubles
TD3 – Triple doubles
PTS – Points
+/- – Plus Minus
FP – Fantasy Points

Figure 2: Glossary for the NBA dataset

Display pairwise correlations like in the graph below. Use both Bonferonni and Holm correction techniques to assess which correlations are truly statistically significant. Explain how these techniques work and what the main differences are between Holm and Bonferonni. What other correction techniques could have been used ?

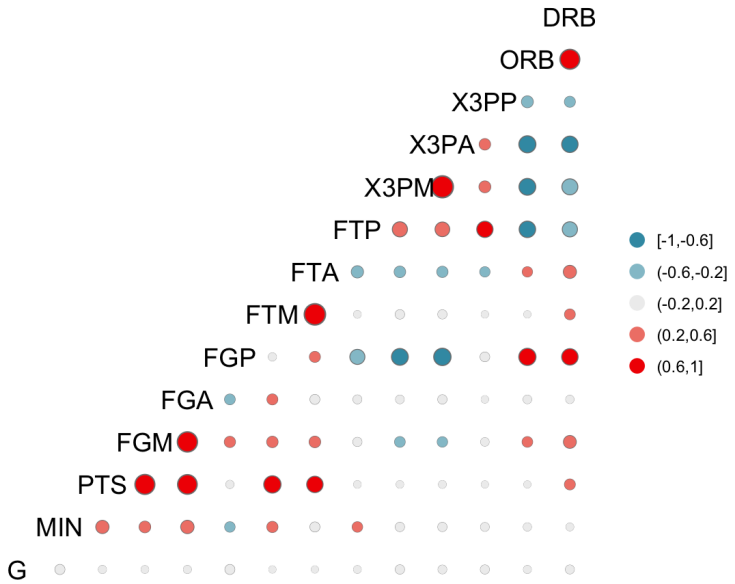

Figure 3: Pairwise correlations in the NBA dataset with ggcor (R)

4 Market Basket Analysis

Create a Jupyter Notebook (Python), and follow the instructions of Datacamp's class on Market Basket Analysis, but apply it on the discipline tag data from theses.fr (in the MBA dataset folder), in addition to the online retail data.

1. In the introduction of the Jupyter Notebook, define precisely the following metrics : support, lift, confidence, leverage and conviction. Plagiarism will not be tolerated, use your own words

2. Explain the notion of pruning, and how the a priori algorithm works with a chart of your own creation (i.e., not found online). Identify frequent itemsets using the apriori method (both using theses tags and retail data)

3. Create visualizations of association rules with plotly when possible. You can find interactive plotly graphs, in the MBA dataset folder, and R codes to produce interactive graphs at this address. You should produce, among other graphs, support / lift scatterplots, parallel plots, and heatmaps. What are the insights that you get from such visualizations ? Elaborate with a 5 to 10 lines long paragraph

enumerate
https://chart-studio.plotly.com/ phoochka/109/association-rules//

5 Exploring sequential data

By contrast with the association-rule mining techniques that we have seen before (like the Apriori algorithm), we are in this case interested in the order in which items appear (or are purchased, in a case of a Market Basket). We will expand the question of sequential pattern mining to the issue of repeated motifs in a sequence. We will begin with n-grams and end with the study of itemsets, like sales and transactions from a e-commerce website.

5.1 Presentation of the dataset

You are provided with an artificial dataset of Tinder conversations.

1. messageid : id of the message
2. conversationid : id of the conversation in which the message can be found (two users per conversation)
3. sender : id of the user who sent the message
4. reader : id of the user who received the message
5. timestamp : timestamp of the message
6. N.words: Number of words in the message
7. rank : Rank of the message in a given conversation. rank = 1 means it is the first message of the conversation
8. rank.inv corresponds to the invert of rank in a given database 1 is the last message of the conversation
9. sentim.ana : sentiment analysis score of the message (it can be negative; the more positive words, the higher the score)
10. message.type : We have built a typology of messages to help you look for patterns. This where we introduced patterns.
    enumerate

    We defined the following mutually exclusive message types :

1. aff.norm : text with normal affirmative sentences, without a question mark, without offensive words, not too many words
2. chunk : text with normal affirmative sentences, without a question mark, without offensive words, but with higher number of words
3. questn : there is a question mark in the text
4. aff.off : text with normal affirmative sentences, without a question mark, not too much text, with offensive words/emojis (insults, strong words, etc).
5. phone : a phone number is exchanged
6. piss.off : one of the users tells the other to stop sending messages, in a not so polite way
7. end : polite message in which a user says that the conversation should stop, without exchanging phone numbers

5.2 Simple motif extraction with the n-gram approach

N-grams can be useful beyond the realm of text mining. Erasmus students who have never followed text mining classes should ask help from fellow Bachelor students. N-grams are notably used in DNA sequencing to extract motifs of interest. Apply this approach, that you have used in previous years, to the messages database and identify most common trigrams, and then to the loan process data. You must to prepare the data (through tokenization, notably) to be able to follow a classic n-grams approach. Which motifs appear to be recurrent ? What metric would you follow to assess their frequency ? Your final output should resemble the table below (except the 1 should not appear). How would you interpret these frequent motifs ?

	Freq <dbl>	Percent <dbl>
questn/3-end/1		
chunk/3-end/1		
chunk/2-questn/1-end/1		
aff.nor/3-end/1		
aff.off/3-end/1		

Figure 4: Representing a subset of ordered sequences of conversation endings

5.3 Sequence data : from description to clustering

Your mission is now to understand how conversations end. Focus solely on the last six messages. Use the R TraMineR package to represent conversation endings (the last 6 messages) without ordering them, like in the figure below.

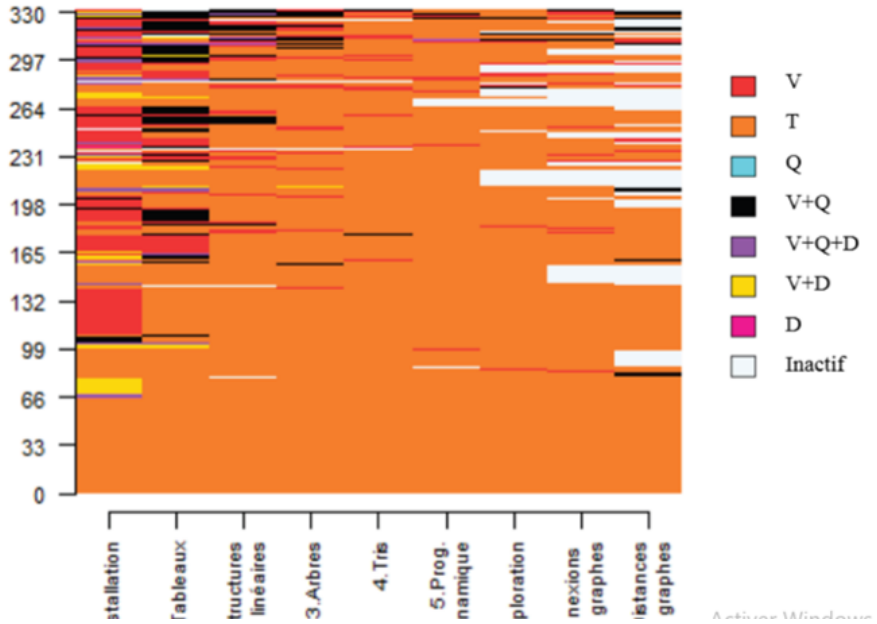

Figure 5: Example of unclustered and unprocessed sequences of states in a MOOC (V = Only viewing videos

Represent a transition matrix for the final two messages of the conversation. A scheme of a state transition matrix is provided in the figure below. Propose an interpretation of what you observe.

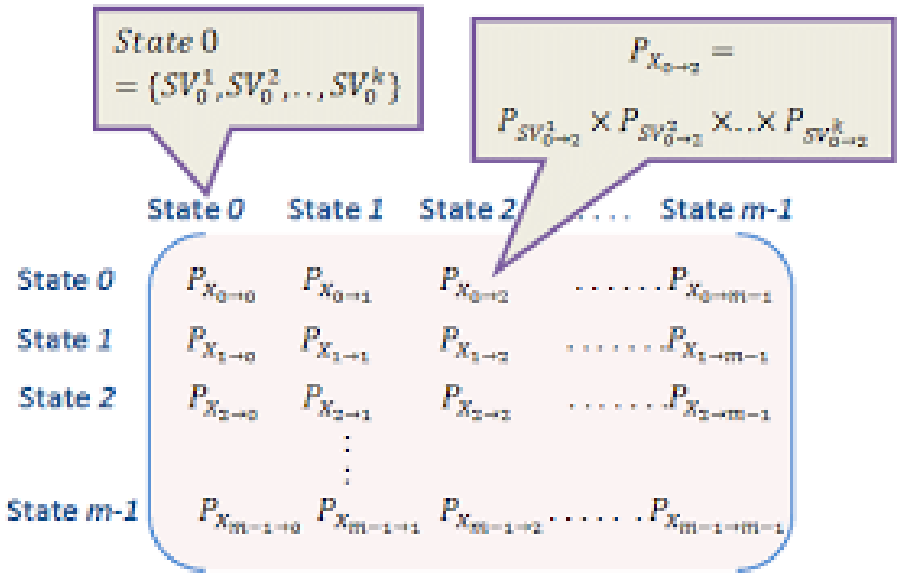

Figure 6: Structure of a state transition matrix

Find the command that allows you to order sequences and produce a graphic similar to the one below.

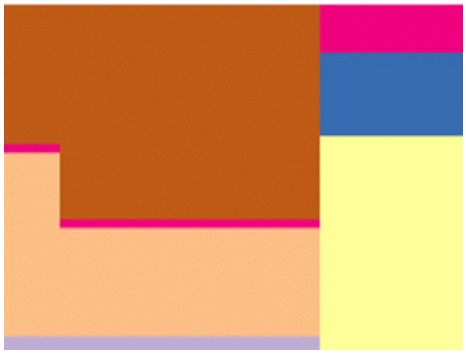

Figure 7: Ordered sequences of conversation endings - colorized

Find a way to separate into four clusters sequences of messages for the last six messages of the conversation. Explain how the clustering algorithm that you used works and provide the formula. How would you label the different groups that you obtained through this clustering method ?

6 Sequential Pattern Mining

In this section, we focus once more on association rules. The transactions that we will study vary in length (by contrast with the previous exercise). In this exercise, you will study the Prefixspan algorithm, as taken from two different github sources. A dataset is provided to you in the Prefixspan data folder.

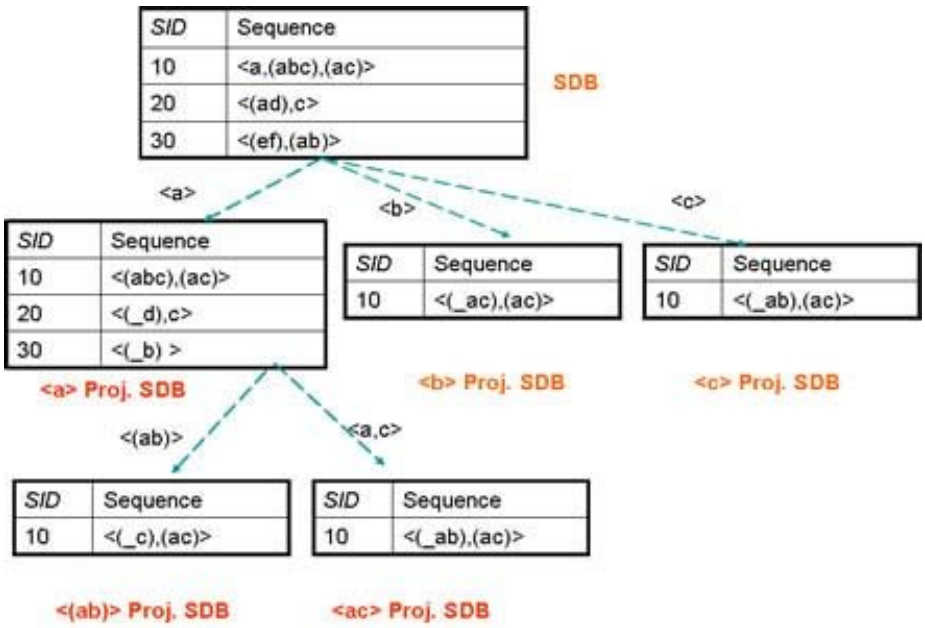

Figure 8: Prefixspan algorithm

1. In around 10 to 20 lines, explain the main differences between GSP, SPADE and Prefixspan algorithms.
2. Apply both GSP and Prefixspan algorithms from the two different Github sources, and play with the parameters (min support, etc.). Display some relevant results. You should reach minsupport=8000 at some point, for GSP.
3. Results differ significantly between Prefixspan and GSP. Are such differences expected ? What could be the causes underlying such differences ?

   enumerate

7 Episode mining

The following section is inspired from the Manila et al. article called "Discovery of Frequent Episodes in Event Sequences".

Episode mining represents a subdomain of process mining. In the figure below, we provide an example of a process as extracted from raw data. Mining can help uncover frequent processes. Abstractly, the data can be viewed as a sequence of events, where each event has an associated time of occurrence, which is an additional data compared to the previous use case. However, there are several areas of application where order between data is important, and, more importantly, timestamps. For example, alarms in a telecommunications network, where thousands of alarms accumulate daily; there can be hundreds of different alarm types. Often the analysis of a large amount of events is quite complicated due to the variability of the actual processes.

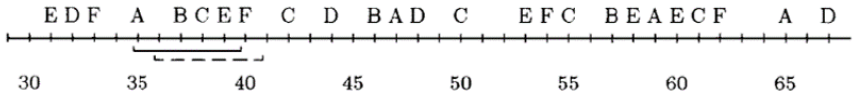

Figure 9: Episode detection and sliding windows

Therefore, it is interesting to search for more local events patterns – find frequent episodes (i.e: collections of events occurring frequently together)– in a large events sequences. For example, when discovering frequent episodes in a telecommunication network alarm log, the goal is to find relationships between alarms. Such relationships can then be used in the on-line analysis of the incoming alarm stream, e.g., to better explain the problems that cause alarms, to suppress redundant alarms, and to predict severe faults.

The goal of this exercise is, given a class of episodes and an input sequence of events, to find all episodes that occur frequently in the event sequence. Several algorithms are proposed to answer this problem.

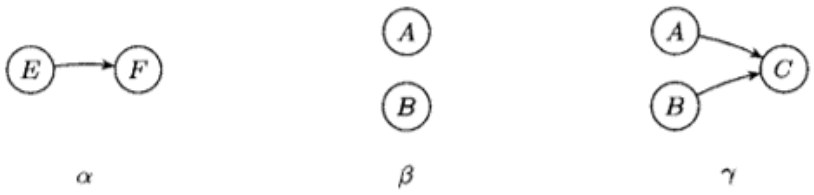

Figure 10: A typology of episodes

In the following exercises, we will focus on episodes, which are subsets of

larger processes, and use winepi and minepi algorithms from a paper of Manila et al. called "Discovery of Frequent Episodes in Event Sequences".

1. Explain the difference between a serial, a parallel episode, non serial non parallel episode involving three different events
2. Explain the differences between the MINEPI and the WINEPI algorithms
3. Apply them to find significant episodes in the loan process dataset that we provided in the WINEPI/MINEPI datasets. Present some relevant results
4. In order for you to get a better grasp on the concept of episode, you must create artificial datasets by groups of 3 or 4 students. In other words, use rules to create episodes in a sequence (using if/then, for instance), share your artificial datasets with other groups. They must discover the rules that you
enumerate

8 Periodic pattern mining

Periodic pattern mining corresponds to the search of sequences of events that are repeated in a string of events (by contrast with itemsets mining). A dataset on quantified self, detailing the actions of a person over the course of several years. A companion code is provided, the goal of this session is to detect periodic patterns in the dataset based on this code. A typical pattern and its different periodicities are displayed in the figure below.

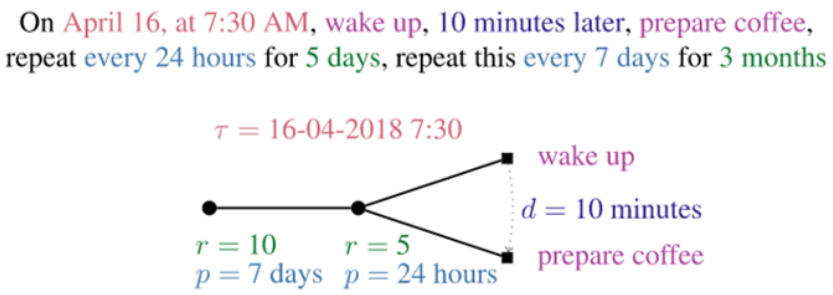

Figure 11: Periodic pattern mining

9 Spatial associations

With regards to spatial associations, we are going to mine for rules in a dataset involving crime data with other characteristics of the environment and of the perpetrator (age, ethnicity, etc.). For instance the figure below features maps representing locations for both crimes and alcoholic drinking places. Data processing for this kind of task is a challenging task. We therefore provided you with a ready-made Jupyter notebook, available at this address, and a dataset, available in the spatial association folder, on Teams, along with the Jupyter notebooks.

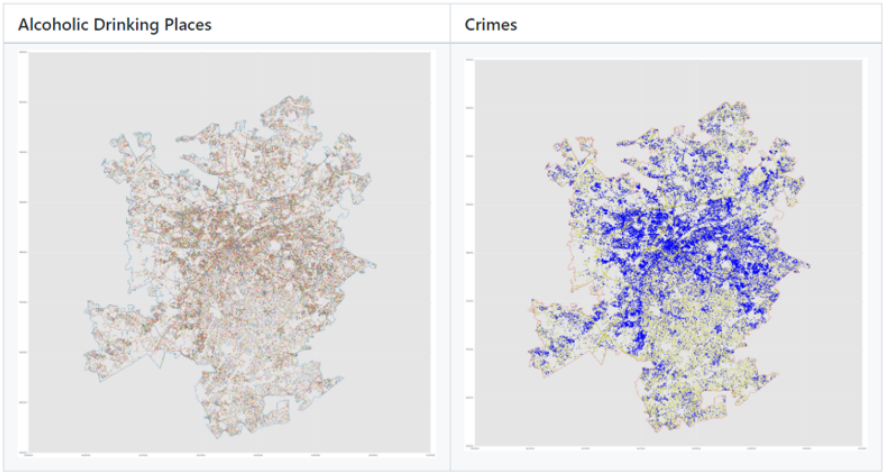

Figure 12: Spatial associations : crime and alcohol selling places

You must implement the code on your own computer. However, you must comment the Jupyter notebook extensively, in order to prove that you have understood the meaning of the different steps that you are following. Your grade, for this section, will depend upon your ability to reproduce the analysis, and on the extent and relevance of your comments. However, refrain from commenting trivial steps.

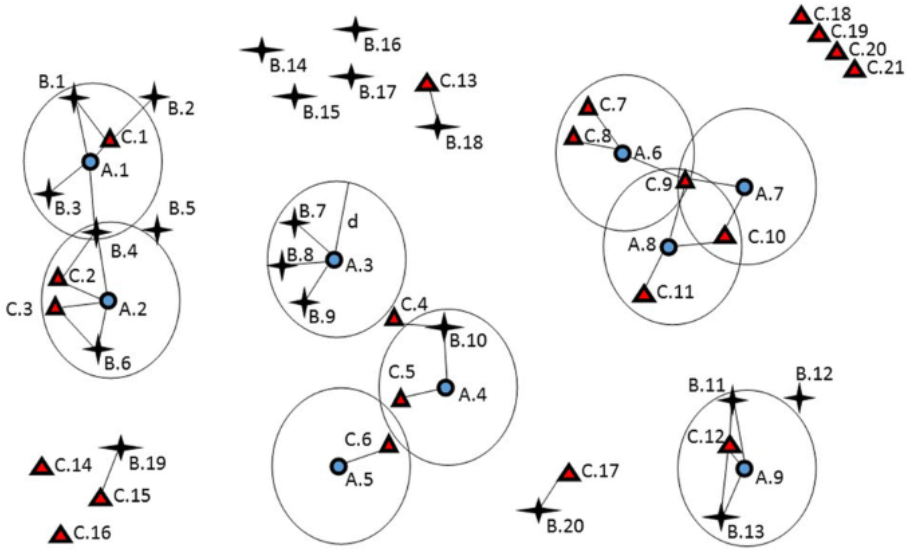

Figure 13: Understanding spatial associations

10 Bonus challenge: quality phrase mining

If you reach and manage to do this part of the project, you will get extra points. A presentation on Autophrase will be made by the instructor. The corresponding code is provided in Teams. We have compiled astracts from a set of thousands of PhD manuscripts into one file per discipline. Your goal is to use quality phrases (Autophrase) the most relevant bigrams and trigrams, and, from this output, to create a word cloud per discipline, with the top 30 "phrases". By contrast with previous sections, where the code is provided and no model training is required, this section implies more autonomous work from your side. We hereby test your ability to autonomously solve technical model training issues.

11 Presentations

By groups of 2 or 3 students, prepare a 30 minutes long presentation (including 5 minutes of questions) exploring both theory and applications for the following pattern mining fields:

1. Frequent subgraph mining - dynamic graph mining
2. Periodic movement mining
3. Semantic rich trajectory mining - Splitter

4. Image mining
5. Flocks, convoys and swarms
6. Phrase mining
7. Shapelet mining

   The presentation will comprise at least twenty slides, whose aim is to cover the most important notions associated with the topic at hand. The corresponding grade will be associated to the Pattern mining learning unit. The notions are important for these courses, and we will rely mostly upon your presentation to teach it. You are responsible for the learning process of the whole classroom, so we have high expectations when it comes to the quality of your explanations.

   We strongly encourage you to use illustrations from textbooks for theoretical topics, and from academic papers (paperswithcode is a great source of free papers with great illustrations) or otherwise when we deal with applications. Also, keep in mind that your presentation should be aesthetically pleasing, and that you must not create cognitive overloads (when it comes to text, less is more, 3 or 4 bulletpoints per slide is a maximum). Each team must choose one topic.

   Half of the topics were inspired by the Pattern discovery in data mining class from Coursera (University of Illinois at Urbana-Champaign). While these courses (mostly weeks 3 and 4) certainly represent a starting point, you must go beyond them. In addition to your slides, you must design five quizzes on the content of your slides. You will post them on Socrative and broadcast them at the end of your presentation in order to make sure that everyone has paid attention to your talk.

   The exact dates of the presentations will be discussed during the classroom sessions.
   enumerate