

CAB330: Case Study 2

Mining from Clothing, Supermarket, News stories and Web Log Data

Due date: 21st Oct 2018
Weighting: 25%

Introduction

The purpose of this assignment is to give you an understanding that data mining methods can be applied to various types of data sets such as record data, transactional data, text data and web logs, and show you the benefits of applying mining techniques to data domains of any kind. This assignment is divided into four parts: Clustering, Association mining, Text Mining and Web Mining. You will use Python with all of the libraries you have learned to use so far

Distribution of Marks (Total: 25 marks)

We would mark your data mining projects in the Week 13 practical class. Every member in the team should complete each of the tasks and should be prepared to show the final models and results panels to your marker. The marker will ask each individual student questions and will assign individual mark (~15%).

In data mining, there is hardly ever a single solution. Also, many times, there is no correct or wrong solution. You may find that your project partner may have different solution as yours. Your group should decide on a single project that you would like to be marked. Submit the report discussing the final project components.

The marks are distributed as follows:

Part 1: Clustering pre-processing and K-means analysis (7 marks)

Part 2: Association Mining and its pre-processing (5 marks)

Part 3: Text Mining (Clustering) the News Stories (7 marks)

Part 4: Web Mining the Log Data for a Website (6 marks)

Part 1: Clustering pre-processing and K-means analysis

MENS_CLOTHING_SALES_2018.csv is a data set contains the StoreCode, annual sales, Sales per square meter and other characteristics of 400 Dutch men's fashion stores. Each row represents an individual store. There are five columns in the data set.

Name	Description
StoreCode	Numerical code for the store (occurs only once in the table)
AnnualSales	Annual sales in Dutch guilders
Sales	Sales per square meter
SFloorSize	Sales floor space of the store (in m ²).
TotalInvestment	Investment in shop-premises and automation.

The purpose of this task is to identify different clusters of men's fashion stores based on the investment and sales. This analysis helps for any new investors to decide investment and location that can return good sales. The analysis can also identify the fashion interests of people from various locations.

Your task is to conduct k-means clustering on this dataset, and find and describe the **minimum number of effective clusters**. Answer the followings in relation to this data and analysis.

Task 1. Data Preparation for Clustering.

1. Can you identify data quality issues in this dataset such as unusual data types, missing values and others?
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. Identify a store that is underperforming in sales. Based on your reporting, the company does not want to focus their efforts on this store. ***Now onwards, the selected store should not be part of analysis.***

Task 2. The first clustering model

1. Build a default clustering model with K= 3 and answer the followings:
 - a. How many records are assigned into each cluster?
 - b. Plot the cluster distribution using pairplot. Explain key characteristics of each cluster/segment.
2. What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better clustering solution?
3. Interpret the (best out of 2.1 and 2.2) cluster analysis outcome. In other words, characterize the nature of each cluster by giving it a descriptive label by using distplot.

Task 3. Refining the clustering model

1. Using elbow method and silhouette, find the optimal K. What is the best K? Explain your reasoning. Evaluate the result.
2. What is the best number of clusters that can describe the dataset effectively? Was this obtained with the default setting (i.e. the automated process) or manually specifying a clustering number?
3. How the outcome of this study can be used by decision makers?

Part 2: Association Mining and its pre-processing

A supermarket store is interested in determining the associations between items purchased from a stationary department and electronics department. The store has chosen to conduct a market basket analysis of specific items purchased from these two departments.

The POS_TRANSACTIONS_2018 data set includes over 400,000 transactions made over the past three months. The following products are represented in the data set:

[A4 copy paper, Drink bottle, Exercise book, USB Flash Drive, DVD media, Sketching Markers, Watercolor Set, Mini Stationery Set, Lanyards, Wristbands, Laminator, Power Bank, Photo Frame, Certificate Frame, Digital Clock, Flash Card, Puzzle]

Name	Description
LOCATION	Point of sale device identification number (e.g. for Register 3)
TRANSACTION_ID	Unique transaction identification number for a given sale. A sale may include several products and thus the same transaction id may occur over several rows.
TRANSACTION_DATE	Date of transaction
PRODUCT_NAME	Product Purchased
QUANTITY	Quantity of this product purchased (always set to 1 by a point of sale device)

Your task is to conduct association analysis on this data set. Answer the followings in relation to this data and analysis.

Task 4. Association Mining

1. Can you identify data quality issues in this dataset for performing association analysis?
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. Conduct association mining and answer the following:
 - a. What is the highest lift value for the resulting rules? Which rule has this value?
 - b. What is the highest confidence value for the resulting rules? Which rule has this value?
 - c. Plot the confidence, lift and support of the resulting rules. Interpret them to discuss the rule-set obtained.
4. The store is particularly interested in products that individuals purchase when they buy “Exercise book”.
 - a. How many rules are in the subset?
 - b. Based on the rules, what are the other products these individuals are most likely to purchase?
5. How the outcome of this study can be used by decision makers?

Part 3: Text Mining (Clustering) the News Stories

Task 5. Text Mining

A leading news corporation, BBC, is planning to start an online personalised news story service. They have a collection of individual stories in the form of a compressed single file. Perform text mining on the BBC(bbc.json) dataset to determine clusters of stories based on similar topics. Answer the following in relation to this data and analysis.

1. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
2. Can you identify data quality issues in order to perform text mining?
3. Based on the ZIPF plot, list the top 10 terms that will be least useful for clustering purpose.
4. Did you disregard any frequent terms? Justify your selection.
5. Justify the term weighting option selected.
6. What is the number of input features available to execute clustering?
(FYI: Note how the original text data is converted into a feature set that can be mined for knowledge discovery.)
7. State how many clusters are generated? Name each cluster meaningfully according to the terms that appear in the clusters?
8. Identify the first fifteen high frequent terms (that are not stop words or noise) in the start list?
9. Describe how these clusters can be useful in the online personalised news story service planned.

Part 4: Web Mining the Log Data for a Website

Task 6. Web Mining

For an e-commerce business, the website structure and site plan were established with the efficiency and usability in mind, but its effectiveness was not verified. Only basic statistics have been produced through simple report and query techniques, but they provide no means for sophisticated web site analysis and predictions. Your task is to determine the patterns of user browsing the website and analyse those patterns to provide the results and recommendations to the website owner.

You have been provided with a log file in CSV format, WEB_LOG_DATA.csv. This was originally a text file and was processed with the steps required for web usage mining as explained in the lecture. The processing steps were: (1) removing unproductive items from the log file such as graphics, sound etc; and (2) identifying users and sessions based on IP address, date and time. The goal of user session identification is to divide the page access of each user into individual sessions.

The dataset consists of 6 columns namely IP address, timestamp, request, step, session id and user id.

Your task is to **apply a data mining operation**, such as classification or clustering or association mining, to the pre-processed data set. Answer the followings in relation to this data and the analyses that you have chosen.

1. For each data mining operation:
 - a. Rationale behind selecting the data mining method.
 - b. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
 - c. Can you identify data quality issues in order to perform web mining?

- d. Discuss the results obtained. Discuss also the applicability of findings of the method. You should include only a high-level managerial kind of discussion on the findings. It should not just be interpretation of results as shown in results panel.

Instructions

1. The assignment is due on 21st Oct at 11:59pm. It is a firm deadline.
2. You should submit the assignment report via Blackboard Assignment.
3. The assignment (record, transaction, text and web mining project DMProj2) will be **marked in the practical class in Week 13**. We will check the outputs along with the assignment report, to assign you marks. The entire team should be present to show the project result and answer the questions raised by marker. We will ask questions to each student, and will assign about 15% of total marks as per individual performance.
4. The datasets required for this assignment can be found on BlackBoard with the file named as **casestudy2-data.zip**. It includes four datasets:
 - a. MENS_CLOTHING_SALES_2018.csv to perform clustering
 - b. POS_TRANSACTIONS_2018.csv to perform association mining
 - c. bbc.json to perform text mining
 - d. WEB_LOG_DATA.csv to perform web mining
5. Name the case-study report as **casestudy2.docx or casestudy2.pdf**. The submission file should include a cover page with Student ID number and full name (as in QUT-Virtual) for all students, along with the group name. Combine this file with your **team contract**, and name the compressed file as **casestudy2.zip**. Submit the zip file on **Blackboard (under assessment panel Assignment 2)**.
6. A report should be submitted via online submission answering each question of the case study. There is no need of including introduction, summary, conclusion or references in the report. The report should just include responses to the questions set in the case-study. Some answers may require screen shots. Use them as needed. You can even include your own table detailing those results based on the outcomes. While you may like to go into extreme detail about, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have thought the matter through.
7. This is a group assignment. The team size is three. You can continue the same group as in case study 1. If you have formed a new group after assignment 1, please notify the lecturing staff. They will remove you from the existing group. In this case, you need to register your new team at Blackboard.
8. The group is to be ARRANGED and MANAGED by you. As in real life, the performance of the individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.
9. Of course, the work your group hand in must be your own; no collaboration or borrowing from others groups is permitted. Read the Assessment Policies on Blackboard or QUT Website.

Assignment Criteria Sheet

Criteria	Comments and scoring
Non Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model with /without submission and demonstrates the ability to run the program and add some components.	1-5
Has demonstrated a task with a working model having a data source, and models with substantial but incorrect implementation of at least one of the four parts. Questions were poorly answered.	6-11
Has implemented all tasks with at least two being substantially correct. Shows some understanding of concepts with some success in applying knowledge. Only basic questions were answered.	12
Has implemented all the tasks: One mining task is fundamentally correct, with substantially correct results which may contain minor errors. Response to questions shows fundamental understanding of terms and concepts.	13-15
Has fundamentally correct implementation of all tasks i.e. selection of correct variables in data, correct allocations, understanding, and explanation of clusters, findings association rules, finding clusters in text data with good term features, and application of an appropriate data mining operation to the web log data. Shows competency in applying text mining. Many questions have been reasonably answered. Demonstrate a good understanding of the methods and terms used in clustering, association mining, text mining and web mining, during written and verbal analyses. Some minor errors are allowed. Written application is required to be of reasonable standard.	16-18
Has implemented all of the requirements above with very few errors. A strong focus on application of tools, and evaluation and interpretation of results is evident.	19-21
All of the criteria above are met, extensive model generation and analyses have been conducted to produce exceptional outcomes. Have applied principles learnt in lectures to enhance the results.	22-25