CS6220 Data Mining Techniques – Spring 2017 Assignment 2

Submission Instructions

- It is recommended that you complete this exercises in **Python 3** and submit your solutions as a **Jupyter notebook**.
- You may use any other language, as long as you **include a README** with simple, clear instructions on how to run (and if necessary compile) your code.
- Please upload all files (code, README, written answers, etc.) to **blackboard** in a single **zip file** named $\{firstname\}\_\{lastname\}\_CS6220\_HW2.zip$.

Understanding Apriori and FP growth

1. Consider a dataset for frequent set mining as in the following table where we have 6 binary features and each row represents a transaction.

TID	Items
1	{c,e}
2	{b,c,d,f}
3	{a,e}
4	{a,b,c}
5	{d}
6	{a,d,f}
7	{c,d,e,f}
8	{a,c,e}
9	{a,d}
10	{b,c,f}

   (a) Illustrate the first three passes of the Apriori algorithm (set sizes 1, 2 and 3) for support threshold of 3 transactions. For each stage, list the candidate sets C_k and the frequent sets L_k. What are the maximal frequent sets discovered in the first 3 levels?

   (b) Pick one of the maximal sets and check if any of its subsets are association rules with frequency at least 0.3 and confidence at least 0.6. Please explain your answer and show your work.

2. Given the following transaction database, let the min_support = 2, answer the following questions.

TID	Items
1	{a,b,e}
2	{a,b,c,d}
3	{a,c,d}
4	{a,c,e}
5	{b,c,f}
6	{a}
7	{a,b,c}
8	{b,d,e}
9	{a,c}
10	{a,b,d,e}

   (a) Construct FP-tree from the transaction database and draw it here.

   (b) Show d's conditional pattern base (projected database), d's conditional FP-tree, and find frequent patterns based on d's conditional FP-tree.

Market Basket Analysis of Academic Communities

In this problem, you will try to apply frequent pattern mining techniques to the real world bibliographic dataset from Aminer (`https://aminer.org/`). One thing worth noting is that you are required consider the whole dataset, instead of running with part of the dataset. You may use any Apriori or FP-growth implementation that is made available in existing libraries. We recommend that you start with Spark (`http://spark.apache.org/`).

To run Spark and Jupyter via Docker:

- Install Docker by following instuctions on

  `https://docs.docker.com/engine/installation/`

- Start up a `jupyter/pyspark` instance

```
docker run -it --rm -p 8888:8888 \
  -v "$PWD":/home/jovyan/work
  jupyter/pyspark-notebook
```

You may also install Spark, Pyspark and Jupyter on your local machine:

- Instructions for Mac OS (using Homebrew)

  `https://www.dataquest.io/blog/pyspark-installation-guide/`

- Instructions for Ubuntu

  `https://roshansanthosh.wordpress.com/2016/02/23/apache-spark-pyspark-standalone-installation-on-ubuntu-14-04/`

1. The dataset included with this problem is `publications.txt`. Parse this data, and comment on how it differs from the previous file (`AP_Train.txt`), in terms of number of publications, authors, venues, references, and years of publication.

2. *Coauthor discovery*: Please use FP-Growth to analyze coauthor relationships, treating each paper as a basket of authors.

   (a) What happens when you successively decrease the support threshold using the values {1e-4, 1e-5, 0.5e-5, 1e-6}?

   (b) Keep threshold = 0.5e-5 and report the top 5 co-authors for these researchers: Rakesh Agrawal, Jiawei Han, Zoubin Ghahramani and Christos Faloutsos according to frequency.

3. *Academic community discovery*: In computer science communities tend to organize around conferences. Here are 5 key conferences for areas of data science

   - Machine learning: NIPS (Neural Information Processing Systems)
   - Data mining: KDD (Conference on Knowledge Discovery and Data Mining)
   - Databases: VLDB (Very Large Data Bases)
   - Computer networks: INFOCOM (International Conference on Computer Communications)
   - Natural language processing: ACL (Association for Computational Linguistics)

   (a) We will now use FP growth to analyze academic communities. To do so, represent each author as a basket in which the items are the venues in which the author has at least one publication. What happens as you decrease the support threshold using values {1e-3, 0.4e-3, 1e-4}?

   (b) Keep the threshold=0.4e-3 and report results. For each area, based on seed conferences please rank the top 10 venues that authors also publish in.