DS5230/DS4420 Unsupervised Machine Learning and Data Mining – Fall 2018 – Homework 1

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 $\{firstname\}\_\{lastname\}\_DS5230/DS4220\_HW1.zip$.

Before you start coding

Please follow the instruction from the attached file called

`Instructions on pySpark.pdf`

in order to either setup your local machine or login onto class server so that you can use jupyter with pyspark for exercise 4.

Exercise 1 : Probability Basics

1. Let X and Y be two independent random variables with densities $p(x)$ and $p(y)$, respectively. Show the following two properties:

$$\mathbb{E}_{p(x,y)}[X + aY] = \mathbb{E}_{p(x)}[X] + a\mathbb{E}_{p(y)}[Y] \tag{1}$$

$$\text{Var}_{p(x,y)}[X + aY] = \text{Var}_{p(x)}[X] + a^2\text{Var}_{p(y)}[Y] \tag{2}$$

for any scalar constant $a \in \mathbb{R}$. Hint: use the definition of expectation and variance,

$$\mathbb{E}_{p(x)}[X] = \int_x p(x)x dx \tag{3}$$

$$\text{var}_{p(x)}[X] = \mathbb{E}_{p(x)}[X^2] - \mathbb{E}^2_{p(x)}[X] \tag{4}$$

2. Let X be a random variable with Beta distribution,

$$p(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \tag{5}$$

where $B(\alpha, \beta)$ is beta function. Prove that

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta} \tag{6}$$

$$\text{var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{7}$$

3. Suppose we observe N i.i.d data points $D = \{x_1, x_2, ..., x_N\}$, where each $x_n \in \{1, 2, ..., K\}$ is a random variable with categorical (discrete) distribution parameterized by $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)$, i.e.,

$$x_n \sim Cat(\theta_1, \theta_2, ..., \theta_K), \quad n = 1, 2, ..., N \tag{8}$$

In detail, this distribution means that for a specific n, the random variable x_n follows $P(x_n = k) = \theta_k$, $k = 1, 2, ..., K$.

Equivalently, we can also write the density function of a categorical distribution as

$$p(x_n) = \prod_{k=1}^{K} \theta_k^{\mathbb{I}[x_n = k]} \tag{9}$$

where $\mathbb{I}[x_n = k]$ is called identity function, and defined as

$$\mathbb{I}[x_n = k] = \begin{cases} 1, & \text{if} \quad x_n = k \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

   a. Now we want to prove that the joint distribution of multiple i.i.d categorical variables is a multinomial distribution. Show that the density function of $D = \{x_1, x_2, ..., x_N\}$ is

$$p(D|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k} \tag{11}$$

   where $N_k = \sum_{n=1}^{N} \mathbb{I}[x_n = k]$ is the number of random variables belonging to category k. In other word, $D = \{x_1, x_2, ..., x_N\}$ follows a multinomial distribution.

   b. We often call $p(D|\boldsymbol{\theta})$ likelihood function, since it indicates the possibility we observe this dataset given the model parameters $\boldsymbol{\theta}$. By Bayes rule, we can rewrite the posterior as

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)} \tag{12}$$

   where $p(\boldsymbol{\theta})$ is piror distribution which indicates our preknowledge about the model parameters. And $p(D)$ is the distribution of the observations (data), which is constant w.r.t. posterior. Thus we can write

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{13}$$

If we assume the Dirichlet prior on $\boldsymbol{\theta}$, i.e.,

$$p(\boldsymbol{\theta}; \alpha_1, \alpha_2, ..., \alpha_K) = Dir(\boldsymbol{\theta}; \alpha_1, \alpha_2, ..., \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{14}$$

where $B(\boldsymbol{\alpha})$ is Beta function and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$.
Now try to derive the joint distribution $p(D, \boldsymbol{\theta})$ and ignore the constant term w.r.t. $\boldsymbol{\alpha}$. Show that the posterior is actually also Dirichlet and parameterized as follows:

$$p(\boldsymbol{\theta}|D) = Dir(\boldsymbol{\theta}; \alpha_1 + N_1, \alpha_2 + N_2, ..., \alpha_K + N_K) \tag{15}$$

[In fact, this nice property is called conjugacy in machine learning. A general statement is : If the prior distribution is conjuagate to the likelihood, then the posterior will be the same distribution as the prior distribution. Search conjugate prior and exponential family for more detail if you are interested.]

Exercise 2 : Exploratory Analysis and Data Visualization

In this exercise, we will be looking at a public citation dataset from Aminer (https://aminer.org/), a free online service used to index and search academic social networks. You will perform some exploratory analysis and data visualization for this dataset. The dataset is up to year 2012 and can be downloaded in the attached file called `q2_dataset.txt`. The data format is documented in the `readme.txt` file. On the server, both dataset and readme file are put in the directory

`/home/yourusername/assignment_dataset/homework1/`.

P.S.: ArnetMiner public citation dataset is a real world dataset containing lots of noise. For example, you may see venue name like "The Truth About Managing People...And Nothing But the Truth". However, you are not required to do data cleaning in this phase.

1. Count the number of distinct authors, publication venues (conferences and journals), and publications in the dataset.

   a. List the each of the counts.

   b. Are these numbers likely to be accurate? As an example look up all the publications venue names associated with the conference "Principles and Practice of Knowledge Discovery in Databases"[1].

   c. In the example above a single venue is listed under multiple names. What additional problem arises when you try to determine the number of distinct authors in a dataset?

[1] https://en.wikipedia.org/wiki/ECML_PKDD

2. We will now look at the publications associated with each author and venue.

    a. For each author, construct the list of publications. Plot a histogram of the number of publications per author (use a logarithmic scale on the y axis).

    b. Calculate the mean and standard deviation of the number of publications per author. Also calculate the Q_1 (1st quartile[2]), Q_2 (2nd quartile, or median) and Q_3 (3rd quartile) values. Compare the median to the mean and explain the difference between the two values based on the standard deviation and the 1st and 3rd quartiles.

    c. Now construct a list of publications for each venue. Plot a histogram of the number of publications per venue. Also calculate the mean, standard deviation, median, Q_1 and Q_3 values. What is the venue with the largest number of publications in the dataset?

3. Now construct the list of references (that is, the cited publications) for each publication. Then in turn use this set to calculate the number of citations for each publication (that is, the number of times a publication is cited).

    a. Plot a histogram of the number of references and citations per publication. What is the publication with the largest number of references? What is the publication with the largest number of citations?

    b. Calculate the so called impact factor for each venue. To do so, calculate the total number of citations for the publications in the venue, and then divide this number by the number of publications for the venue. Plot a histogram of the results.

    c. What is the venue with the highest apparent impact factor? Do you believe this number?

    d. Now repeat the calculation from item b., but restrict the calculation to venues with at least 10 publications. How does your histogram change? List the citation counts for all publications from the venue with the highest impact factor. How does the impact factor (mean number of citations) compare to the median number of citations?

    e. Finally, construct a list of publications for each publication year. Use this list to plot the average number of references and average number of citations per publication as a function of time. Explain the differences you see in the trends.

Exercise 3 : 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.

[2] https://en.wikipedia.org/wiki/Quartile

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 datasbase 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.

Exercise 4 : 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/).

If you use the class server to do exercise 4:

==========

Due to some mismatch of python version between the master node and worker node in pyspark, you would encounter an error when initialing a pyspark object,i.e., using the method *pyspark.sparkContext*(). The current solution to this issue I found so far, is to add the following two lines of code to your script before you initialize pyspark object :

```
import os
os.environ['PYSPARK_PYTHON'] = '/opt/anaconda3/bin/python3'
```

I have tried all kinds of ways to adding env variable to the PATH, but unfortunately it still cannot be solved on system level. I am very sorry for the inconvenience.

==========

1. The dataset included with this problem is **q4_dataset.txt**. Parse this data, and comment on how it differs from the previous file (**q2_dataset.txt**), in terms of number of publications, authors, venues, references, and years of publication.On the server, both dataset and readme file are put in the directory

   `/home/yourusername/assignment_dataset/homework1/`.

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.