DS 5230 Unsupervised Machine Learning and Data Mining
Homework 1 By Yanchi Li

Exercise 1

Problem 1 a. According to the definition of Marginal Probability Density Function, we have

$$f_X(x) = \int f_X(x, y)dy \tag{1}$$

$$f_Y(y) = \int f_Y(x, y)dx \tag{2}$$

Thus,

$$\begin{aligned}
E_{p(x,y)}[X + aY] &= \int\int_{p(x,y)} (x + ay)dxdy \\
&= \int\int_{p(x,y)} xdxdy + \int\int_{p(x,y)} aydxdy \\
&= \int_{p(x)} xdx + \int_{p(y)} aydy \\
&= E_{p(x)}[X] + aE_{p(y)}[Y]
\end{aligned} \tag{3}$$

b. Since X and Y are independent,

$$E_{p(x,y)}[XY] = E_{p(x,y)}[X]E_{p(x,y)}[Y] \tag{4}$$

Then using the results from part 1, we can get

$$\begin{aligned}
Var_{p(x,y)}[X + aY] &= E_{p(x,y)}[X^2 + 2aXY + Y^2] - E^2_{p(x,y)}[X + aY] \\
&= E_{p(x,y)}[X^2 + 2aXY + Y^2] - (E^2_{p(x,y)}[X] + a^2E^2_{p(x,y)}[Y] \\
&\quad + 2aE_{p(x,y)}[X]E_{p(x,y)}[Y]) \\
&= E_{p(x,y)}[X^2] - E^2_{p(x,y)}[X] + E_{p(x,y)}[Y^2] - E^2_{p(x,y)}[Y] \\
&\quad + 2aE_{p(x,y)}[XY] - 2aE_{p(x,y)}[X]E_{p(x,y)}[Y]) \\
&= Var_{p(x,y)}[X] + aVar_{p(x,y)}[Y]
\end{aligned} \tag{5}$$

Problem 2 a. By definition, we have

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1} = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{6}$$

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \tag{7}$$

Then the proof is quite straightforward

$$
\begin{aligned}
E[X] &= \int_0^1 \frac{x^{\alpha-1} \cdot (1-x)^{\beta-1}}{B(\alpha,\beta)} dx \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+1) \cdot \Gamma(\beta) \cdot \Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1) \cdot \Gamma(\beta) \cdot \Gamma(\alpha)} \\
&= \frac{\alpha\Gamma(\alpha) \cdot \Gamma(\beta) \cdot \Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta) \cdot \Gamma(\beta) \cdot \Gamma(\alpha)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
\tag{8}
$$

b.

$$
\begin{aligned}
Var(x) &= E[X^2] - E^2[X] \\
&= \int_0^1 x^2 p(x) dx - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \int_0^1 \frac{x^{\alpha+1} \cdot (1-x)^{\beta-1}}{B(\alpha,\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\Gamma(\alpha+2) \cdot \Gamma(\beta) \cdot \Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+2) \cdot \Gamma(\beta) \cdot \Gamma(\alpha)} - \frac{\alpha^2}{(\alpha+\beta)^2} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha(\alpha+1)\Gamma(\alpha) \cdot \Gamma(\alpha+\beta) \cdot \Gamma(\beta)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta) \cdot \Gamma(\alpha) \cdot \Gamma(\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}
\end{aligned}
\tag{9}
$$

Problem 3 a. Since we have the density function of a categorical distribution which is

$$
p(x_n) = \prod_{k=1}^{K} \theta_k^{\mathbb{I}[x_n=k]} \tag{10}
$$

and all those categorical distributions are independent, so we just multiply the density function of them to get the joint distribution.

$$
\begin{aligned}
p(D|\boldsymbol{\theta}) &= \prod_{n=1}^{N} p(x_n) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_k^{\mathbb{I}[x_n=k]} \\
&= \prod_{k=1}^{K} \theta_k^{\sum_{n=1}^{N} \mathbb{I}[x_n=k]} \\
&= \prod_{k=1}^{K} \theta_k^{N_k}
\end{aligned} \tag{11}
$$

b. Since the distribution of the observations $p(D)$ is constant with respect to the posterior. So when ignoring the constant term, we can rewrite posterior by Bayes rules as

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

By the conclusion of previous part, we have

$$
p(D|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k} \tag{13}
$$

and we have already assume the Dirichlet prior on $\boldsymbol{\theta}$ is

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

where $B(\alpha)$ is Beta function and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$.

So we simply multiply those two equations to derive the joint distribution $p(D, \boldsymbol{\theta})$ which is also posterior function since the constant term has been ignored.

$$
\begin{aligned}
p(\boldsymbol{\theta}|D) &= p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \prod_{k=1}^{K} \theta_k^{N_k} \cdot \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\
&= \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \\
&= Dir(\boldsymbol{\theta}; \alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)
\end{aligned} \tag{15}
$$

Exercise 3

Problem 1

a. In words, the A-priori algorithm can be divided into two parts: self-joining and pruning. In each pass, we first construct a candidate set which includes all itemsets that fits the target size and the minimum support threshold. The trick here is we only find pair of sets in L_{k-1} that differ by exactly one element. Then, by removing all candidates with infrequent subsets we can get a frequent itemsets of target size.

First pass:
We start by counting all those size-1 frequent itemsets, which are: Those size-1

Table 1: Candidate Sets and Frequent Sets C_1 and L_1

Items	Frequent Itemsets of Size 1
{a}	5
{b}	3
{c}	6
{d}	5
{e}	4
{f}	4

frequent itemsets serve both as the candidate sets C_1 and the frequent sets L_1.

Second pass:
Since all size-1 itemsets are frequent, the candidate sets C_2 is the full permutation of all six items. We also count the frequency of each candidate sets which are shown as Table 2.

Table 2: Candidate Sets C_2

Candidate Sets C_2	Frequency	Candidate Sets C_2	Frequency
{a, b}	1	{b, f}	2
{a, c}	2	{c, d}	2
{a, d}	2	{c, e}	2
{a, e}	2	{c, f}	3
{a, f}	1	{d, e}	1
{b, c}	3	{d, f}	3
{b, d}	1	{e, f}	1
{b, e}	0		

According to the instructions, the support threshold is set to be 3 transactions, so all candidate sets with frequency less than 3 would not be considered as frequent itemset. Pruning those sets would results in the frequent sets of size-2 L_2 shown in Table 3.

Table 3: Frequent Sets L_2

Frequent Sets	Frequency
{b, c}	3
{c, f}	3
{d, f}	3

Third pass:

The only three frequent sets of size-2 are {b, c}, {c, f}, {d, f}, so there are also only three candidate sets for size of 3, which are:

Table 4: Candidate Sets C_3

Candidate Sets	Frequency
{b, c, f}	2
{c, d, f}	2
{b, d, f}	1

Since all the candidate sets have frequency less than 3, there is no frequent sets for size of 3 and any other larger sizes. We may say the maximal frequent sets are {b, c}, {c, f}, {d, f} because they are frequent and there is no frequent superset of them.

b. Let's say, pick {c, f} to check association rules with the subsets of it. In particular, we want to see whether people are likely to buy item c when they already buy item f. The support is

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N} = \frac{3}{10} \tag{16}$$

and the confidence is

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{X} = \frac{3}{4} \tag{17}$$

Problem 2

a. According to the given transaction database, we can construct a FP-tree which is shown below.

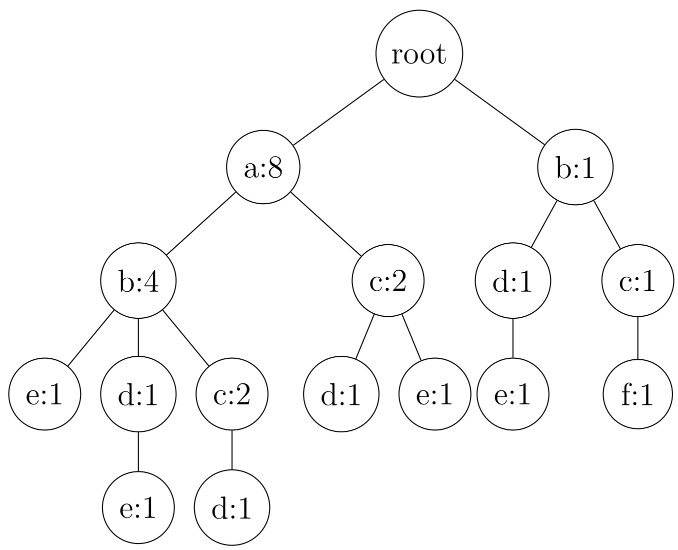

Figure 1: FP-tree

b. First we find d's sub-tree and d's conditional FP-tree.

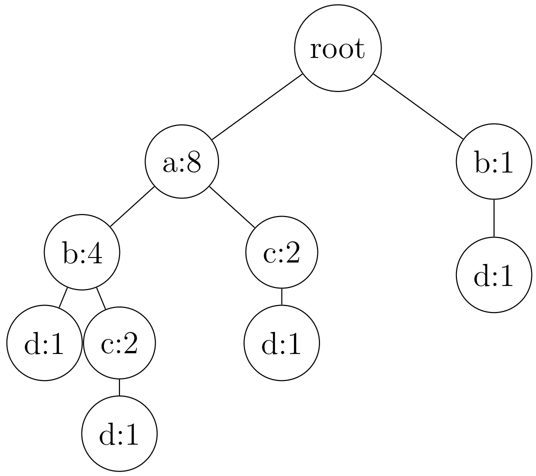

Figure 2: d's Sub-tree

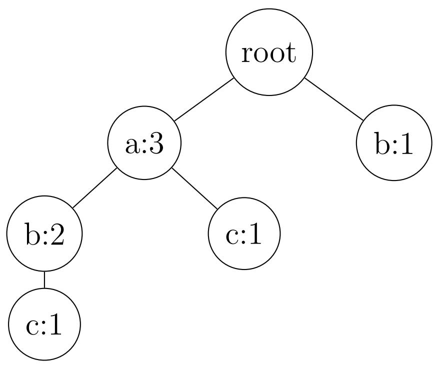

Figure 3: d's Conditional Sub-tree

Then we will get d's conditional pattern base from d's conditional sub-tree, which is

Table 5: d's Conditional Pattern Base

Item	Conditional Pattern Base
d	ab: 1, abc: 1, ac: 1, b: 1

Finally, we may find frequent patterns of $\{a, d\}$, $\{c, d\}$ and $\{b, d\}$ based on d's conditional FP-tree.