Exercise A3

Statistical Analysis of Big Data

Amanda Magzal 207608647

(a) Number of Possible Rules

Suppose there are d items in a data set. Let A denote the items that form the left hand side of the rule, and B the items that form the right hand side of the rule, creating a rule $A \to B$.

First, we choose $k < d$ items to form A. There are $\binom{d}{k}$ ways to do this. Then, we choose i items from the remaining $d-k$ items to form B. There are $\binom{d-k}{i}$ ways to do this.

The total number of rules R is:

$$R = \sum_{k=1}^{d} \binom{d}{k} \sum_{i=1}^{d-k} \binom{d-k}{i}$$

Since

$$\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1$$

$$R = \sum_{k=1}^{d} \binom{d}{k} (2^{d-k} - 1)$$

$$= \sum_{k=1}^{d} \binom{d}{k} 2^{d-k} - \sum_{k=1}^{d} \binom{d}{k}$$

$$= \sum_{k=1}^{d} \binom{d}{k} 2^{d-k} - (2^d + 1)$$

Since

$$(1+x)^d = \sum_{i=1}^{d} \binom{d}{i} x^{d-i} + x^d$$

substituting $x = 2$ leads to:

$$3^d = \sum_{i=1}^{d} \binom{d}{i} 2^{d-i} + 2^d$$

Therefore,

$$R = 3^d - 2^d - (2^d + 1)$$

$$= 3^d - 2^{d+1} + 1 \blacksquare$$

(b) Association Rules

Suppose X is a frequent itemset. All rules generated from X satisfy the support threshold, as each of their support is identical to the support for X.

The confidence of a rule $A \rightarrow B$ generated from X is:

$$c(A \rightarrow B) = \frac{\sigma(X)}{\sigma(A)}$$

Where $\sigma(I)$ denotes the support count of itemset I.

Consider the following transaction data:

Table 1: Transaction Data

TID	Beer	Eggs	Flour	Milk
1	0	1	1	1
2	1	1	1	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1

In order to find all rules with $minsup = 0.4$, we first find frequent itemsets with support count ≥ 2.

Table 2: 1-itemsets

(a) Candidates

Items	Support Count
{Beer}	1
{Eggs}	4
{Flour}	3
{Milk}	4

(b) Frequents

Items	Support Count
{Eggs}	4
{Flour}	3
{Milk}	4

Table 3: 2-itemsets

(a) Candidates

Items	Support Count
{Eggs, Flour}	3
{Eggs, Milk}	3
{Flour, Milk}	2

(b) Frequents

Items	Support Count
{Eggs, Flour}	3
{Eggs, Milk}	3
{Flour, Milk}	2

Table 4: 3-itemsets

(a) Candidates

Items	Support Count
{Eggs, Flour, Milk}	2

(b) Frequents

Items	Support Count
{Eggs, Flour, Milk}	2

Next, we generate all possible rules from the frequent itemsets. For each rule we calculate the support and confidence.

For example, for the rule $\{\text{Eggs}\} \rightarrow \{\text{Flour}\}$, the support is

$$s\big(\{\text{Eggs}\} \rightarrow \{\text{Flour}\}\big) = \frac{\sigma\big(\{\text{Eggs, Flour}\}\big)}{N} = \frac{3}{5} = 0.6$$

and the confidence is

$$c\big(\{\text{Eggs}\} \rightarrow \{\text{Flour}\}\big) = \frac{\sigma\big(\{\text{Eggs, Flour}\}\big)}{\sigma\big(\{\text{Eggs}\}\big)} = \frac{3}{4} = 0.75$$

Table 5 shows all possible rules with their corresponding support and confidence.

Table 5: Possible Association Rules

Rule	Support	Confidence
{Eggs} → {Flour}	0.6	0.75
{Flour} → {Eggs}	0.6	1
{Eggs} → {Milk}	0.6	0.75
{Milk} → {Eggs}	0.6	0.75
{Flour} → {Milk}	0.4	0.67
{Milk} → {Flour}	0.4	0.5
{Eggs} → {Flour, Milk}	0.4	0.5
{Milk} → {Eggs, Flour}	0.4	0.5
{Flour} → {Eggs, Milk}	0.4	0.67
{Eggs, Flour} → {Milk}	0.4	0.67
{Flour, Milk} → {Eggs}	0.4	1
{Eggs, Milk} → {Flour}	0.4	0.67

All rules satisfy the required $minsup = 0.4$ as they were generated from frequent itemsets. Any rules that do not satisfy $minconf = 0.7$ should be eliminated. The final rules are shown in table 6.

Table 6: Association Rules

Rule	Support	Confidence
{Eggs} → {Flour}	0.6	0.75
{Flour} → {Eggs}	0.6	1
{Eggs} → {Milk}	0.6	0.75
{Milk} → {Eggs}	0.6	0.75
{Flour, Milk} → {Eggs}	0.4	1

(c) Statistical Model

Let $X = (X_1, X_2, ..., X_k)$ be a K-dimensional random vector of possibly correlated Bernoulli random variables (binary outcomes), and let $x = (x_1, x_2, ...x_k)$ be a realization of X. The joint probability density is of the form:

$$p(x) = p_{0,0,...0}^{[\prod_{j=1}^{k}(1-x_j)]} \; p_{1,0,...,0}^{[x_1 \prod_{j=2}^{k}(1-x_j)]} \; p_{0,1,...,0}^{[(1-x_1)x_2 \prod_{j=3}^{k}](1-x_j)]} \; ... p_{1,1,...1}^{[\prod_{j=1}^{k} x_j]}$$

The multivariate Bernoulli can be used to formulate the graph structure of binary variables.

A graph model considers a graph $G = (V, E)$, whose nodes set V represents k random variables $X_1, X_2, ...X_k$ connected or disconnected defined by the undirected edges set E. This formulation allows pairwise relationships among the nodes to be described in terms of edges, which in statistics are defined as correlations.

Transactions of market basket data can be considered realizations of a multivariate Bernoulli distribution with k items. The item has value 1 if it was purchased in the transaction, and value 0 otherwise.

Combined with the graph model, the multivariate Bernoulli can be used to estimate pairwise and higher order interctions (i.e. association rules) between the items.

The article used for this question can be found here.

(d) Pseudo-Code to Calculate Support

Algorithm 1: Support Count with MapReduce

```
1  Function Map(id, transaction):
2      for item i in transaction do
3          EmitIntermediate(i, "1")
4      end
5
6  Function Reduce(item, counts):
7      result = 0
8      for value in counts do
9          result += value
10     end
11     return result
```