**Data Mining and Knowledge Discovery (KSE525)**

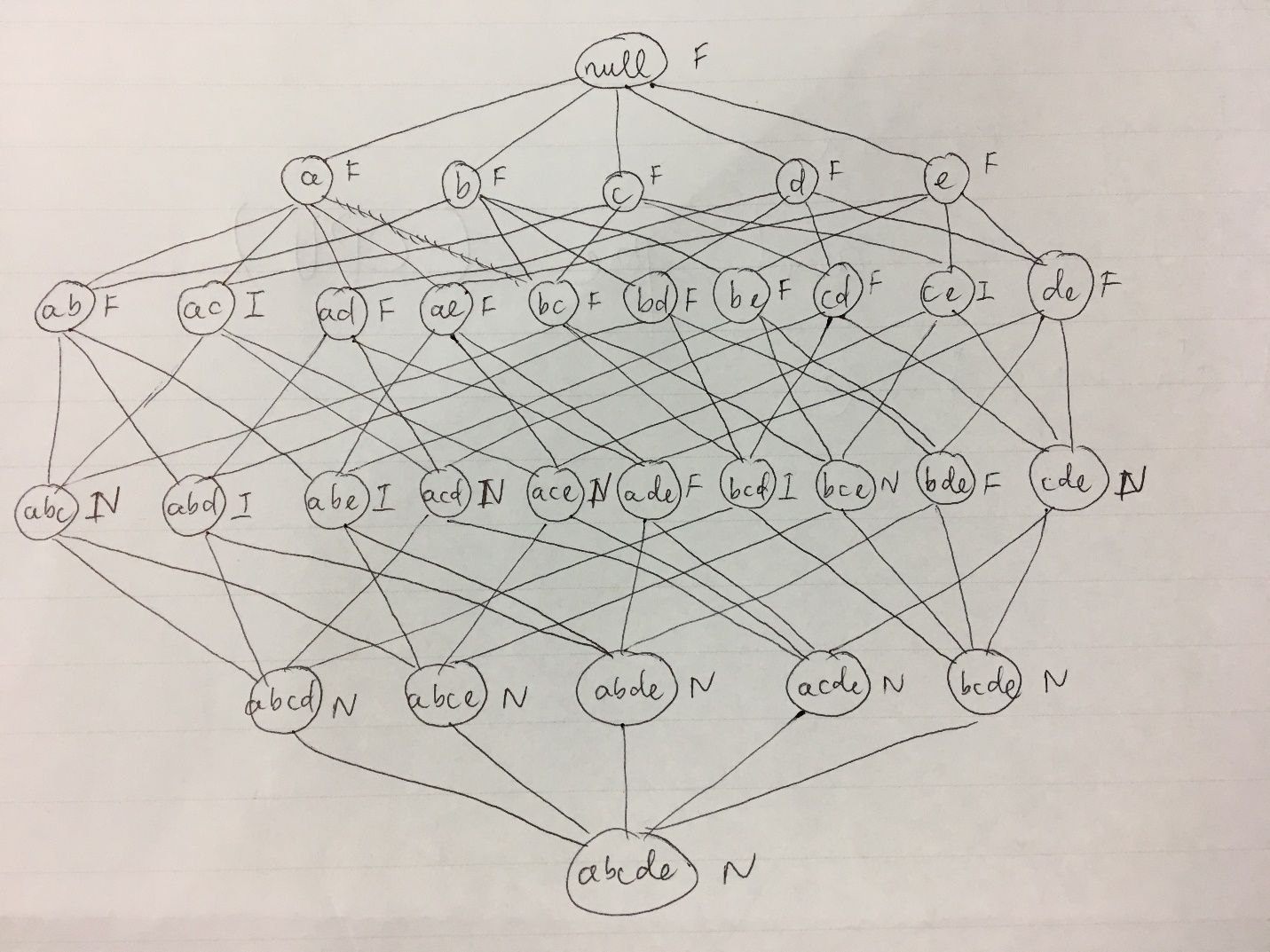
Assignment 2

**Name: Nguyen Ngoc Quang**

**Student ID: 20174556**

1.

A, The itemset lattice representing the dataset:



B, Percentage of frequent itemsets = 16/32 = 50%

C, Pruning ratio of the Apriori algorithm = 11/32 = 34.375%

D, False alarm rate = 5/32 = 15.625%

2.

A, “hot dogs” => “hamburgers” is a strong rule because:

Support = 2000/5000 = 0.4 > minsup = 0.2

Confidence = 2000/3000 = 0.67 > minconf = 0.5

B,

The probability of purchasing hot dogs: p ({hot dogs}) = 3000/5000 = 0.6

The probability of purchasing hamburgers: p ({hamburgers}) = 2500/5000 = 0.5

The probability of purchasing both hot dogs and hamburgers: p ({hot dogs, hamburgers}) = 2000/5000 = 0.4

Lift = p ({hot dogs, hamburgers}) / (p ({hot dogs}) x p ({hamburgers})) = 0.4/ (0.6x0.5) = 1.33 > 1

* The purchase of hot dogs is dependent of the purchase of hamburgers, and the correlation relationship is positively correlated.

C,

Contingency table with expected values:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hamburgers |  |  |
| Hot dogs | 2000 (1500) | 500 (1000) | 2500 |
|  | 1000 (1500) | 1500 (1000) | 2500 |
|  | 3000 | 2000 | 5000 |

Let’s set x: # purchase of hamburgers, y: #purchase of hot dogs, and so on, as the below table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hamburgers |  |  |
| Hot dogs | xy | y | y |
|  |  |  |  |
|  | x |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Set | xy | y |  |  |  | lift | all\_conf. | max\_conf. | Kulc. | cosine |
| D1 | 2000 | 1000 | 500 | 1500 | 833.3 | 1.33 | 0.67 | 0.8 | 0.733 | 0.730 |

The results of all\_conf, max\_conf, Kulc, cosine show that Hot dogs and Hamburgers are positively associated.

Lift > 1, indicates that Hot dogs and Hamburgers are positively associated, while >1, the observed value of (Hot dogs and Hamburgers) is 2000 > the expected value of 1500 indicates that they are positively associated.

The use of these 4 measures are the same as the use of lift and . But still, lift and generate dramatically different measure value due to their sensitivity to , so it may generate unstable results.