Data Mining: Practical 2

YiKai Wang 15206086

Q1:

1): Use data.drop() remove the useless column.

2&3): from the tutorial we know that here we should transform all different categories under all tags into a True-False matrix to suit the algorithm.

The out put result contain 20 frequent item sets (with min_support=0.15) and the support is in [0.2,0.44], maximum is 0.44.

4&5) from the tutorial we know that we could use :

mlxtend.frequent_patterns.association_rules()

to workout the association rules, set metric = 'confidence' and min_threshold=0.9 to suit the requirement of question.

There are only one rule here:

frozenset({'21...25'}),frozenset({'junior'}),
0.16,0.44,0.16,1.0,2.272727272727273,0.0896,inf

Which I think means: 'age in 21-25 years old take the junior lesson.'

6&7) if we set min_threshold to 0.7 we gain 3 results:

a) frozenset({'21...25'}),frozenset({'junior'}),
0.16,0.44,0.16,1.0,2.272727272727273,0.0896,inf

b) frozenset({'Ph.D'}),frozenset({'26...30'}),
0.2,0.32,0.16,0.7999999999999999,2.4999999999999996,0.096,3.3999999999999986

c) frozenset({'philosophy'}),frozenset({'26...30'}),
0.28,0.32,0.2,0.7142857142857143,2.232142857142857,0.1104,2.38

For b: PhD student always in age 26-30

For c: Philosophy students always in age 26-30

Q2:

1&2&3&4)Steps:

1. same as Q1.1, remove ID.
2. Extract numeric data, use np.cut(data, 3) divide them into 3 bins, categorify it.
3. Extract binary data, add the tag name into it, or it will be regard as one category, influence the result.
    a. Codes here:

```
data['married'] = ['Married_'+str(item) for item in data['married']]
data['car'] = ['Car_' + str(item) for item in data['car']]
data['save_act'] = ['Save_Account_' + str(item) for item in data['save_act']]
data['current_act'] = ['Current_Account_' + str(item) for item in
data['current_act']]
data['mortgage'] = ['Mortgage_' + str(item) for item in data['mortgage']]
data['pep'] = ['Pep_' + str(item) for item in data['pep']]
```

4. Use the:

mlxtend.frequent_patterns.fpgrowth()

to workout the frequent itemset. Totally with 231 set (minimum frequency = 0.2) and the biggest one is 0.29 (support).

5&6)

The minimum threshold of confidence is 0.79 and gained 11 items. If we set 0.791-0.793 it will be 10

7.

Interest 1:

((50.667, 67.0]) (Save_Account_YES)

Support = 0.251667 confidence = 0.790576

age in 50-67 always have a save account.

((-0.003, 1.0], Pep_NO) (Married_YES)

support = 0.260000 confidence = 0.812500

the people with one child and no PEP always married