ECT_HW2

2020

(a) 題目

- 請嘗試著修改 adult.csv 的欄位與上圖相同，並轉換成 arff 檔使其可以執行 Association Rule，請說明使用的方法以及解釋原來的檔案不能執行的原因？(10%)

(a) 解答

- 透過文字編輯器修改成如圖所示

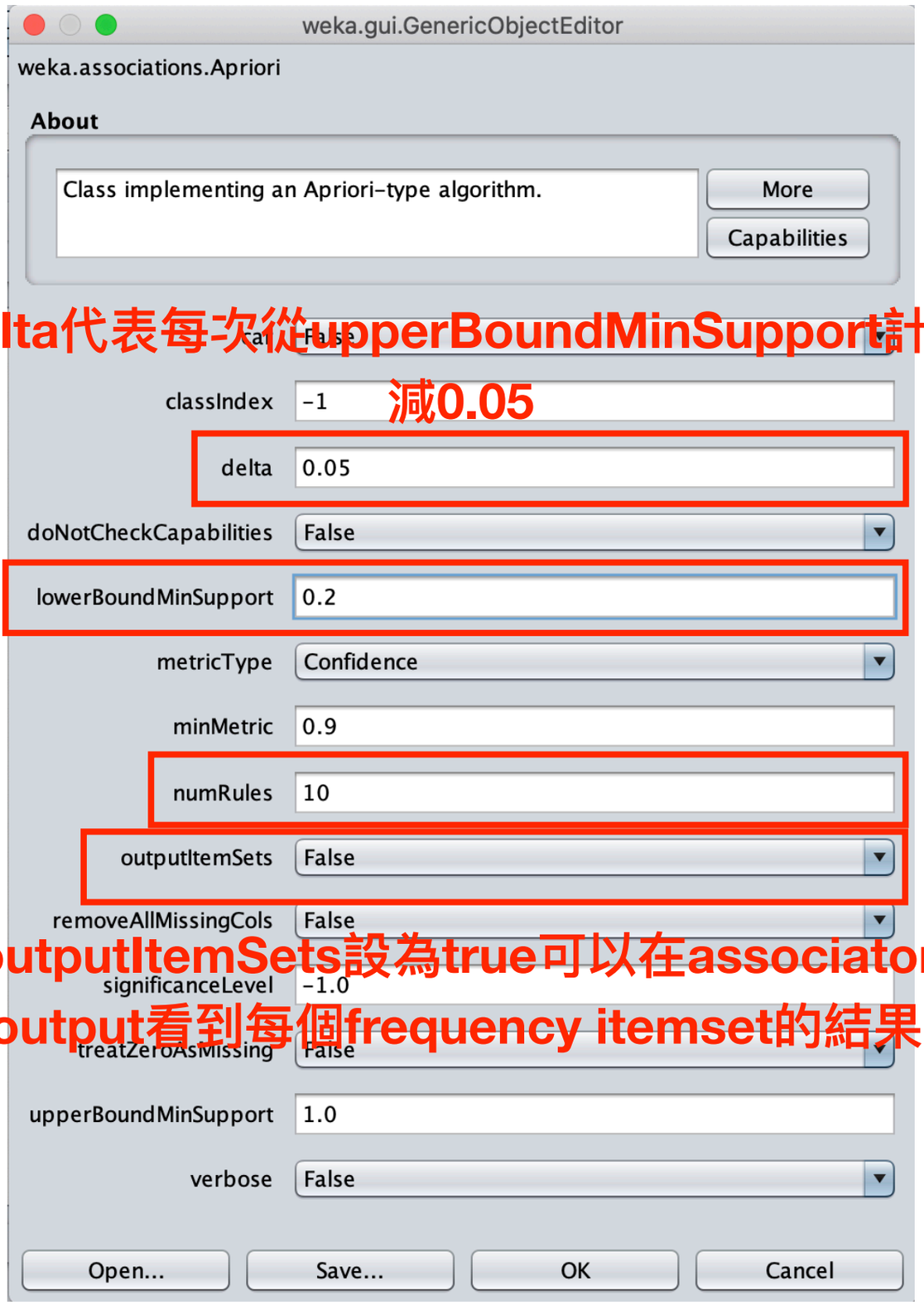
- 在Weka利用前處理 NumericToNominal的方法將Numeric欄位轉換成Nominal

- 原因：Aprioir演算法要求其處理的資料欄位皆為Nominal

(b) 題目

- 請將 numRule 設成 5 和 10，其各別執行後的 Minimum support 為何，請比較兩者並說明造成其差異的原因。(15%)

(b) 答案




```
Apriori                                  Apriori
=======                                  =======

Minimum support: 0.2 (9207 instances)    Minimum support: 0.25 (11508 instances)
Minimum metric <confidence>: 0.9         Minimum metric <confidence>: 0.9
Number of cycles performed: 16           Number of cycles performed: 15
```

- numRules設為10的Minimum support 值0.2，小於numRules設為5的Minimum support值0.25。
- 原因：想要找尋的rule數較多，必須放寬每次篩選通過的數目，所以Minimum support 數值才會比較低，使找到的規則更容易進入下一階段的篩選，最後找到的rule 總數也會比較多；反之當只要5條rules，則篩選通過的數目不需要那麼多，門檻就可以拉高。Minimum support數值才會相對高。

(c) 題目

- 將 numRule 設成 10，列出前 5 條 rule (15%)

(c) 題目

```
Best rules found:

 1. marital-status=Never-married hours-per-week=20-40 9669 ==> income=<=50K 9368    <conf:(0.97)> lift:(1.29) lev:(0.05) [2098] conv:(7.94)
 2. workclass=Private marital-status=Never-married 12243 ==> income=<=50K 11755    <conf:(0.96)> lift:(1.28) lev:(0.06) [2549] conv:(6.21)
 3. workclass=Private marital-status=Never-married race=White 10134 ==> income=<=50K 9702    <conf:(0.96)> lift:(1.27) lev:(0.05) [2082] conv:(5.81)
 4. marital-status=Never-married 14875 ==> income=<=50K 14153    <conf:(0.95)> lift:(1.27) lev:(0.06) [2968] conv:(5.1)
 5. marital-status=Never-married race=White 12228 ==> income=<=50K 11590    <conf:(0.95)> lift:(1.26) lev:(0.05) [2396] conv:(4.75)
 6. gender=Male hours-per-week=40-60 10122 ==> race=White 9388    <conf:(0.93)> lift:(1.08) lev:(0.02) [714] conv:(1.97)
 7. hours-per-week=40-60 12403 ==> race=White 11366    <conf:(0.92)> lift:(1.07) lev:(0.02) [738] conv:(1.71)
 8. age=20-30 11487 ==> income=<=50K 10513    <conf:(0.92)> lift:(1.22) lev:(0.04) [1876] conv:(2.92)
 9. income=>50K 11422 ==> race=White 10367    <conf:(0.91)> lift:(1.06) lev:(0.01) [579] conv:(1.55)
10. marital-status=Married-civ-spouse race=White income=<=50K 10343 ==> gender=Male 9378    <conf:(0.91)> lift:(1.34) lev:(0.05) [2387] conv:(3.47)
```

- 5條rule:

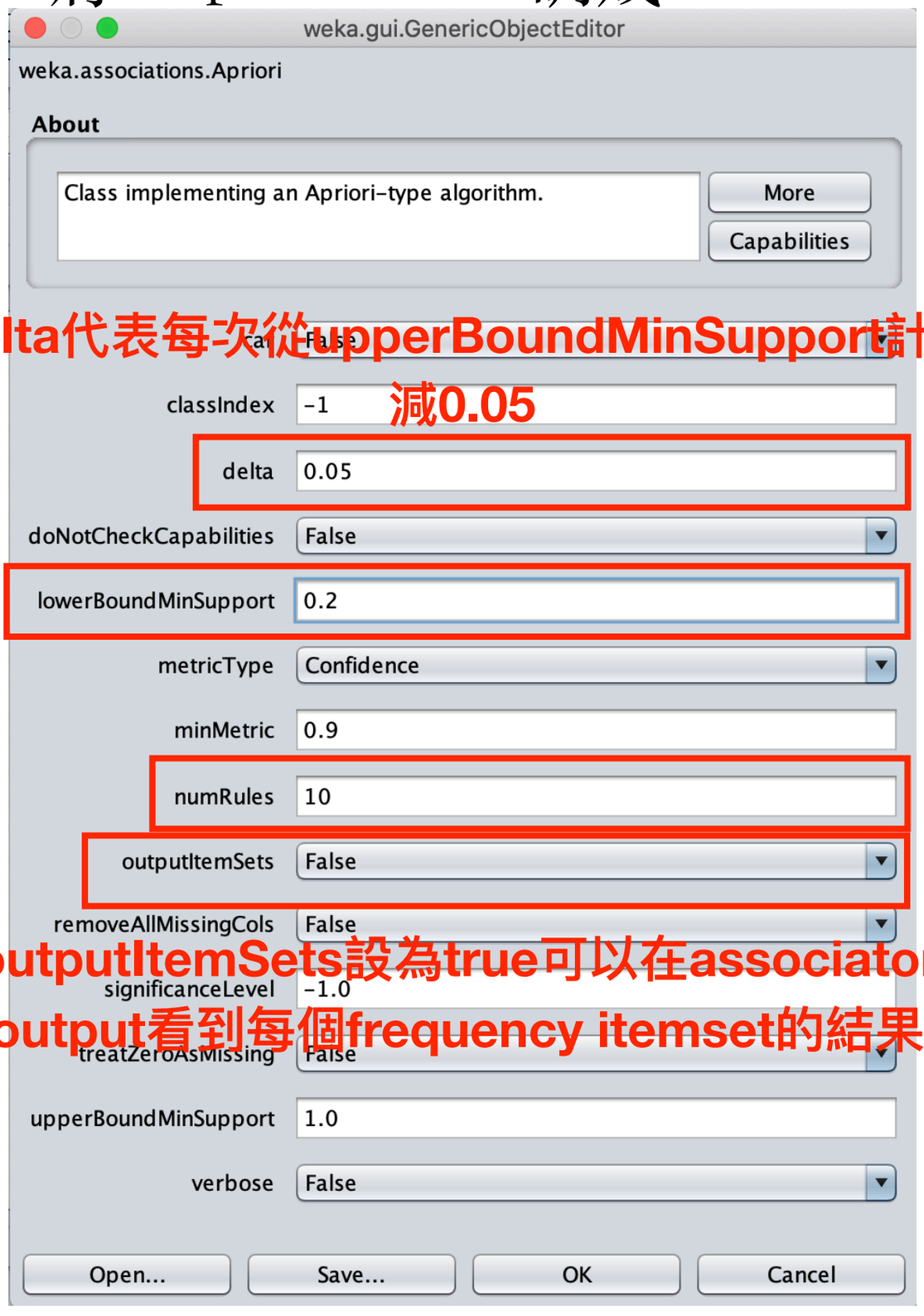
1. If married-status=Never-married and hours-per-week=20-40 then income <=50K
2. If workclass=Private and married-status=Never-married then income<=50K
3. If workclass=Private and married-status=Never-married and race=white then income<=50K
4. If married-status=Never-married then income<=50K
5. If married-status=Never-married and race=white then income<=50K

(d) 題目

- 如何在 Associator output 產生 Itemset，請截圖說明並附上 Itemset 結果。(15%)

(d) 答案

- 將outputItemSets調成true



delta代表每次從upperBoundMinSupport計算減0.05

outputItemSets設為true可以在associator output看到每個frequency itemset的結果

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Large Itemsets L(1):
age=20-30 11487
age=30-40 12538
age=40-50 10182
workclass=Private 33906
education=HS-grad 14972
education=Some-college 10036
marital-status=Never-married 14875
marital-status=Married-civ-spouse 21451
race=White 39444
gender=Male 31114
gender=Female 14919
hours-per-week=20-40 28350
hours-per-week=40-60 12403
income=<=50K 34611
income=>50K 11422
```

(d) 答案

```
Size of set of large itemsets L(2): 38

Large Itemsets L(2):
age=20-30 workclass=Private 9649
age=20-30 race=White 9650
age=20-30 income=<=50K 10513
age=30-40 workclass=Private 9370
age=30-40 race=White 10636
workclass=Private education=HS-grad 11682
workclass=Private marital-status=Never-married 12243
workclass=Private marital-status=Married-civ-spouse 14473
workclass=Private race=White 29024
workclass=Private gender=Male 22307
workclass=Private gender=Female 11599
workclass=Private hours-per-week=20-40 21656
workclass=Private income=<=50K 26519
education=HS-grad race=White 12737
education=HS-grad gender=Male 10251
education=HS-grad hours-per-week=20-40 10123
education=HS-grad income=<=50K 12535
marital-status=Never-married race=White 12228
marital-status=Never-married hours-per-week=20-40 9669
marital-status=Never-married income=<=50K 14153
marital-status=Married-civ-spouse race=White 19229
marital-status=Married-civ-spouse gender=Male 19183
marital-status=Married-civ-spouse hours-per-week=20-40 12062
marital-status=Married-civ-spouse income=<=50K 11705
marital-status=Married-civ-spouse income=>50K 9746
race=White gender=Male 27421
race=White gender=Female 12023
race=White hours-per-week=20-40 23465
race=White hours-per-week=40-60 11366
race=White income=<=50K 29077
race=White income=>50K 10367
gender=Male hours-per-week=20-40 17943
gender=Male hours-per-week=40-60 10122
gender=Male income=<=50K 21386
gender=Male income=>50K 9728
gender=Female hours-per-week=20-40 10407
gender=Female income=<=50K 13225
hours-per-week=20-40 income=<=50K 22833
```

```
Size of set of large itemsets L(3): 29

Large Itemsets L(3):
workclass=Private education=HS-grad race=White 9907
workclass=Private education=HS-grad income=<=50K 9983
workclass=Private marital-status=Never-married race=White 10134
workclass=Private marital-status=Never-married income=<=50K 11755
workclass=Private marital-status=Married-civ-spouse race=White 12941
workclass=Private marital-status=Married-civ-spouse gender=Male 12878
workclass=Private race=White gender=Male 19602
workclass=Private race=White gender=Female 9422
workclass=Private race=White hours-per-week=20-40 17985
workclass=Private race=White income=<=50K 22282
workclass=Private gender=Male hours-per-week=20-40 13422
workclass=Private gender=Male income=<=50K 16015
workclass=Private gender=Female income=<=50K 10504
workclass=Private hours-per-week=20-40 income=<=50K 18043
education=HS-grad race=White income=<=50K 10500
marital-status=Never-married race=White income=<=50K 11590
marital-status=Never-married hours-per-week=20-40 income=<=50K 9368
marital-status=Married-civ-spouse race=White gender=Male 17345
marital-status=Married-civ-spouse race=White hours-per-week=20-40 10483
marital-status=Married-civ-spouse race=White income=<=50K 10343
marital-status=Married-civ-spouse gender=Male hours-per-week=20-40 10482
marital-status=Married-civ-spouse gender=Male income=<=50K 10487
race=White gender=Male hours-per-week=20-40 15331
race=White gender=Male hours-per-week=40-60 9388
race=White gender=Male income=<=50K 18529
race=White gender=Female income=<=50K 10548
race=White hours-per-week=20-40 income=<=50K 18607
gender=Male hours-per-week=20-40 income=<=50K 13450
gender=Female hours-per-week=20-40 income=<=50K 9383
```

```
Size of set of large itemsets L(4): 8

Large Itemsets L(4):
workclass=Private marital-status=Never-married race=White income=<=50K 9702
workclass=Private marital-status=Married-civ-spouse race=White gender=Male 11625
workclass=Private race=White gender=Male hours-per-week=20-40 11463
workclass=Private race=White gender=Male income=<=50K 13829
workclass=Private race=White hours-per-week=20-40 income=<=50K 14774
workclass=Private gender=Male hours-per-week=20-40 income=<=50K 10479
marital-status=Married-civ-spouse race=White gender=Male income=<=50K 9378
race=White gender=Male hours-per-week=20-40 income=<=50K 11345
```

(e) 題目

- 使用已修改過的 adult.csv 檔，使用 Apriori 演算法進行分析,設定 confidence = 0.9、minimum support = 0.2，過程中對所有重要程式步驟進行截圖並加以說明，越詳盡越好。(15%)

(e) 答案

- 讀取資料集

```
import pandas as pd
df = pd.read_csv('adult.csv')
```

- 由於apriori要求數據格式為list，因此要轉換資料格式

```
data = df.values.tolist()
```

(e) 答案

- 引用Apriori，並設定參數

In [26]:
```
from apyori import apriori

rules = list(apriori(data, min_support= 0.2, min_confidence= 0.9))
rules
```

Out[26]:
```
[RelationRecord(items=frozenset({'<=50K', '20-30'}), support=0.22837964069254665, ordered_statistics=[OrderedStatis
tic(items_base=frozenset({'20-30'}), items_add=frozenset({'<=50K'}), confidence=0.9152084965613302, lift=1.21723708
42277807)]),
 RelationRecord(items=frozenset({'40-60', 'White'}), support=0.2469098255599244, ordered_statistics=[OrderedStatist
ic(items_base=frozenset({'40-60'}), items_add=frozenset({'White'}), confidence=0.9163911956784649, lift=1.069471552
3442546)]),
 RelationRecord(items=frozenset({'<=50K', 'Never-married'}), support=0.3074533486846393, ordered_statistics=[Ordere
dStatistic(items_base=frozenset({'Never-married'}), items_add=frozenset({'<=50K'}), confidence=0.9514621848739496,
lift=1.2654548772443017)]),
 RelationRecord(items=frozenset({'White', '>50K'}), support=0.22520800295440227, ordered_statistics=[OrderedStatist
ic(items_base=frozenset({'>50K'}), items_add=frozenset({'White'}), confidence=0.90763438977412, lift=1.059251948698
7138)]),
 RelationRecord(items=frozenset({'<=50K', '20-40', 'Female'}), support=0.20383203354115526, ordered_statistics=[Ord
eredStatistic(items_base=frozenset({'20-40', 'Female'}), items_add=frozenset({'<=50K'}), confidence=0.9016046891515
327, lift=1.1991438749447432)]),
 RelationRecord(items=frozenset({'<=50K', '20-40', 'Never-married'}), support=0.20350618034888016, ordered_statisti
cs=[OrderedStatistic(items_base=frozenset({'20-40', 'Never-married'}), items_add=frozenset({'<=50K'}), confidence=0
.9688695832040543, lift=1.2886069031126588)]),
 RelationRecord(items=frozenset({'40-60', 'Male', 'White'}), support=0.20394065127191363, ordered_statistics=[Order
```

(f) 題目

- 產生與 (c) 小題一樣的結果，列出前五條best rules，截圖並加以說明(15%)

(f) 答案

```
 1. marital-status=Never-married hours-per-week=20-40 9669 ==> income=<=50K 9368    <conf:(0.97)> lift:(1.29) lev:(0.05) [2098] conv:(7.94)
 2. workclass=Private marital-status=Never-married 12243 ==> income=<=50K 11755    <conf:(0.96)> lift:(1.28) lev:(0.06) [2549] conv:(6.21)
 3. workclass=Private marital-status=Never-married race=White 10134 ==> income=<=50K 9702    <conf:(0.96)> lift:(1.27) lev:(0.05) [2082] conv:(5.81)
 4. marital-status=Never-married 14875 ==> income=<=50K 14153    <conf:(0.95)> lift:(1.27) lev:(0.06) [2968] conv:(5.1)
 5. marital-status=Never-married race=White 12228 ==> income=<=50K 11590    <conf:(0.95)> lift:(1.26) lev:(0.05) [2396] conv:(4.75)
 6. gender=Male hours-per-week=40-60 10122 ==> race=White 9388    <conf:(0.93)> lift:(1.08) lev:(0.02) [714] conv:(1.97)
 7. hours-per-week=40-60 12403 ==> race=White 11366    <conf:(0.92)> lift:(1.07) lev:(0.02) [738] conv:(1.71)
 8. age=20-30 11487 ==> income=<=50K 10513    <conf:(0.92)> lift:(1.22) lev:(0.04) [1876] conv:(2.92)
 9. income=>50K 11422 ==> race=White 10367    <conf:(0.91)> lift:(1.06) lev:(0.01) [579] conv:(1.55)
10. marital-status=Married-civ-spouse race=White income=<=50K 10343 ==> gender=Male 9378    <conf:(0.91)> lift:(1.34) lev:(0.05) [2387] conv:(3.47)
```

- Best rules 是以confidence排序

- 整理出rule, support, confidence

```
result = pd.DataFrame()
for item in rules:
    series = pd.Series({"Rule":item[0],"Support":item[1],"Confidence":item[2][0][2]})
    result = result.append(series, ignore_index=True)
```

- 以confidence作排序

```
result.sort_values(by= ['Confidence'], ascending=False)
```

	Confidence	Rule	Support
5	0.968870	(<=50K, 20-40, Never-married)	0.203506
8	0.960140	(<=50K, Private, Never-married)	0.255360
12	0.957371	(<=50K, Private, Never-married, White)	0.210762
2	0.951462	(<=50K, Never-married)	0.307453
9	0.947825	(<=50K, Never-married, White)	0.251776