Data Mining Project 1

資工碩一 P76084300 施逢怡

Dataset: Adult Data Set(https://archive.ics.uci.edu/ml/datasets/Adult)

- Attributes: (age, workclass, education, marital-status, occupation, relationship, race, sex, hrs-per-work, native-country, salary)

Result:

- **Apriori 實作結果**

(support, confidence) = (0.7,0.7)

```
Support Of X: 85.42  Support of X & Y: 79  Confidence: 92  [' White']------>[' United-States']
Support Of X: 89.58  Support of X & Y: 79  Confidence: 88  [' United-States']------>[' White']
```

(support, confidence) = (0.5,0.5)

```
Support Of X: 85.42  Support of X & Y: 59  Confidence: 69  [' White']------>[' Male']
Support Of X: 66.92  Support of X & Y: 59  Confidence: 88  [' Male']------>[' White']
Support Of X: 85.42  Support of X & Y: 79  Confidence: 92  [' White']------>[' United-States']
Support Of X: 89.58  Support of X & Y: 79  Confidence: 88  [' United-States']------>[' White']
Support Of X: 85.42  Support of X & Y: 64  Confidence: 74  [' White']------>[' <=50K']
Support Of X: 75.92  Support of X & Y: 64  Confidence: 84  [' <=50K']------>[' White']
Support Of X: 85.42  Support of X & Y: 60  Confidence: 70  [' White']------>[' Private']
Support Of X: 69.7  Support of X & Y: 60  Confidence: 85  [' Private']------>[' White']
Support Of X: 89.58  Support of X & Y: 60  Confidence: 67  [' United-States']------>[' Male']
Support Of X: 66.92  Support of X & Y: 60  Confidence: 89  [' Male']------>[' United-States']
Support Of X: 89.58  Support of X & Y: 68  Confidence: 75  [' United-States']------>[' <=50K']
Support Of X: 75.92  Support of X & Y: 68  Confidence: 89  [' <=50K']------>[' United-States']
Support Of X: 89.58  Support of X & Y: 62  Confidence: 69  [' United-States']------>[' Private']
Support Of X: 69.7  Support of X & Y: 62  Confidence: 89  [' Private']------>[' United-States']
Support Of X: 69.7  Support of X & Y: 54  Confidence: 78  [' Private']------>[' <=50K']
Support Of X: 75.92  Support of X & Y: 54  Confidence: 72  [' <=50K']------>[' Private']
Support Of X: 85.42  Support of X & Y: 54  Confidence: 63  [' White']------>[' Male', ' United-States']
Support Of X: 89.58  Support of X & Y: 54  Confidence: 61  [' United-States']------>[' Male', ' White']
Support Of X: 66.92  Support of X & Y: 54  Confidence: 81  [' Male']------>[' United-States', ' White']
Support Of X: 78.68  Support of X & Y: 54  Confidence: 69  [' United-States', ' White']------>[' Male']
Support Of X: 58.88  Support of X & Y: 54  Confidence: 92  [' Male', ' White']------>[' United-States']
Support Of X: 59.85  Support of X & Y: 54  Confidence: 91  [' United-States', ' Male']------>[' White']
Support Of X: 85.42  Support of X & Y: 58  Confidence: 68  [' White']------>[' <=50K', ' United-States']
Support Of X: 89.58  Support of X & Y: 58  Confidence: 65  [' United-States']------>[' <=50K', ' White']
Support Of X: 75.92  Support of X & Y: 58  Confidence: 77  [' <=50K']------>[' United-States', ' White']
Support Of X: 78.68  Support of X & Y: 58  Confidence: 74  [' United-States', ' White']------>[' <=50K']
Support Of X: 63.57  Support of X & Y: 58  Confidence: 91  [' <=50K', ' White']------>[' United-States']
Support Of X: 67.56  Support of X & Y: 58  Confidence: 86  [' <=50K', ' United-States']------>[' White']
Support Of X: 85.42  Support of X & Y: 54  Confidence: 64  [' White']------>[' Private', ' United-States']
Support Of X: 89.58  Support of X & Y: 54  Confidence: 61  [' United-States']------>[' Private', ' White']
Support Of X: 69.7  Support of X & Y: 54  Confidence: 78  [' Private']------>[' United-States', ' White']
Support Of X: 78.68  Support of X & Y: 54  Confidence: 69  [' United-States', ' White']------>[' Private']
Support Of X: 59.59  Support of X & Y: 54  Confidence: 91  [' Private', ' White']------>[' United-States']
```

- **Apriori (by Weka)**

//每次最多顯示 5 筆

(1) high support, high confidence (0.7, 0.7)

```
Apriori
=======

Minimum support: 0.7 (22793 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 1

Best rules found:

1. race= White 27816 ==> native-country= United-States 25621    <conf:(0.92)> lift:(1.03) lev:(0.02) [701] conv:(1.32)
2. native-country= United-States 29170 ==> race= White 25621    <conf:(0.88)> lift:(1.03) lev:(0.02) [701] conv:(1.2)
```

(2) high support, low confidence(0.7, 0.1)

```
Apriori
=======

Minimum support: 0.7 (22793 instances)
Minimum metric <confidence>: 0.1
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 1

Best rules found:

 1. race= White 27816 ==> native-country= United-States 25621    <conf:(0.92)> lift:(1.03) lev:(0.02) [701] conv:(1.32)
 2. native-country= United-States 29170 ==> race= White 25621    <conf:(0.88)> lift:(1.03) lev:(0.02) [701] conv:(1.2)
```

(3) Low support, low confidence(0.1, 0.1)

```
Apriori
=======

Minimum support: 0.6 (19537 instances)
Minimum metric <confidence>: 0.1
Number of cycles performed: 8

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 4

Best rules found:

1. race= White 27816 ==> native-country= United-States 25621    <conf:(0.92)> lift:(1.03) lev:(0.02) [701] conv:(1.32)
2. salary= <=50K 24720 ==> native-country= United-States 21999    <conf:(0.89)> lift:(0.99) lev:(-0) [-146] conv:(0.95)
3. workclass= Private 22696 ==> native-country= United-States 20135    <conf:(0.89)> lift:(0.99) lev:(-0.01) [-197] conv:(0.92)
4. native-country= United-States 29170 ==> race= White 25621    <conf:(0.88)> lift:(1.03) lev:(0.02) [701] conv:(1.2)
5. salary= <=50K 24720 ==> race= White 20699    <conf:(0.84)> lift:(0.98) lev:(-0.01) [-418] conv:(0.9)
```

(4) Low support, high confidence(0.1, 0.9)

```
Apriori
=======

Minimum support: 0.5 (16280 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 8

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

Best rules found:

1. race= White 27816 ==> native-country= United-States 25621    <conf:(0.92)> lift:(1.03) lev:(0.02) [701] conv:(1.32)
2. race= White sex= Male 19174 ==> native-country= United-States 17653    <conf:(0.92)> lift:(1.03) lev:(0.01) [475] conv:(1.31)
3. race= White salary= <=50K 20699 ==> native-country= United-States 18917    <conf:(0.91)> lift:(1.02) lev:(0.01) [373] conv:(1.21)
4. workclass= Private race= White 19404 ==> native-country= United-States 17728    <conf:(0.91)> lift:(1.02) lev:(0.01) [344] conv:(1.21)
5. sex= Male native-country= United-States 19488 ==> race= White 17653    <conf:(0.91)> lift:(1.06) lev:(0.03) [1004] conv:(1.55)
```

- **Apriori 比較與心得:**

這次使用了美國人關於職業、教育、性別、種族等的資料集，希望能透過關聯法則找出各個資料是否有關連。

透過實作和 Weka，我從(support, confidence)=(0.9,0.9)開始往下測試，發現 support 值太高，使得程式或是 Weka 都找不出關聯法則，直到 support 值調到 0.7 之後，找到 2 個關聯:United Stated->white 和 white->United Stated，代表的意義是在這個資料集裡面，國籍是美國的人通常是白人，且白人通常是住在美國。

當繼續調低 support 值之後，因為在每個階段中，進入下一個階段的關聯組合太多，除了因為計算量太大而使的程式執行速度太久，得出的關聯法則量也大量成長。

觀察其他的關聯之後，發現出來的法則往往都互相關聯(A->B 且 B->A)，而且感覺都環繞著幾個特徵(美國人，男人，白人，在私人企業工作)。

- **FPGrowth 實作結果**

Support:0.8

```
Support Of X: 85.42  Support of X & Y: 79  Confidence: 92  [' White']------>[' United-States']
Support Of X: 89.58  Support of X & Y: 79  Confidence: 88  [' United-States']------>[' White']
```

Support:0.5

```
Rule 1: native-country: United-States, Support: 0.896
Rule 2: race: White, Support: 0.854
Rule 3: native-country: United-States => race: White, Support: 0.787
Rule 4: salary: <=50K, Support: 0.759
Rule 5: workclass: Private, Support: 0.697
Rule 6: native-country: United-States => salary: <=50K, Support: 0.676
Rule 7: sex: Male, Support: 0.669
Rule 8: race: White => salary: <=50K, Support: 0.636
Rule 9: native-country: United-States => workclass: Private, Support: 0.618
Rule 10: native-country: United-States => sex: Male, Support: 0.598
Rule 11: race: White => workclass: Private, Support: 0.596
Rule 12: race: White => sex: Male, Support: 0.589
Rule 13: native-country: United-States => race: White => salary: <=50K, Support: 0.581
Rule 14: salary: <=50K => workclass: Private, Support: 0.545
Rule 15: native-country: United-States => race: White => workclass: Private, Support: 0.544
Rule 16: native-country: United-States => race: White => sex: Male, Support: 0.542
```

FPGrowth

//每次最多顯示 5 筆

(1) high support, high confidence(0.7, 0.7)

```
FPGrowth found 2 rules (displaying top 2)

1. [race= White_binarized=1]: 27816 ==> [native-country= United-States_binarized=1]: 25621   <conf:(0.92)> lift:(1.03) lev:(0.02) conv:(1.32)
2. [native-country= United-States_binarized=1]: 29170 ==> [race= White_binarized=1]: 25621   <conf:(0.88)> lift:(1.03) lev:(0.02) conv:(1.2)
```

(2) high support, low confidence(0.7, 0.2)

```
FPGrowth found 2 rules (displaying top 2)

1. [race= White_binarized=1]: 27816 ==> [native-country= United-States_binarized=1]: 25621   <conf:(0.92)> lift:(1.03) lev:(0.02) conv:(1.32)
2. [native-country= United-States_binarized=1]: 29170 ==> [race= White_binarized=1]: 25621   <conf:(0.88)> lift:(1.03) lev:(0.02) conv:(1.2)
```

(3) Low support, low confidence(0.1, 0.1)

```
FPGrowth found 6 rules (displaying top 5)

1. [race= White_binarized=1]: 27816 ==> [native-country= United-States_binarized=1]: 25621   <conf:(0.92)> lift:(1.03) lev:(0.02) conv:(1.32)
2. [workclass= Private_binarized=1]: 22696 ==> [native-country= United-States_binarized=1]: 20135   <conf:(0.89)> lift:(0.99) lev:(-0.01) conv:(0.92)
3. [native-country= United-States_binarized=1]: 29170 ==> [race= White_binarized=1]: 25621   <conf:(0.88)> lift:(1.03) lev:(0.02) conv:(1.2)
4. [workclass= Private_binarized=1]: 22696 ==> [race= White_binarized=1]: 19404   <conf:(0.85)> lift:(1) lev:(0) conv:(1)
5. [race= White_binarized=1]: 27816 ==> [workclass= Private_binarized=1]: 19404   <conf:(0.7)> lift:(1) lev:(0) conv:(1)
```

(4) Low support, high confidence(0.2, 0.9)

```
FPGrowth found 11 rules (displaying top 5)

1. [relationship= Husband_binarized=1]: 13193 ==> [marital-status= Married-civ-spouse_binarized=1]: 13184   <conf:(1)> lift:(2.17) lev:(0.22) conv:(712.51)
2. [race= White_binarized=1, relationship= Husband_binarized=1]: 11940 ==> [marital-status= Married-civ-spouse_binarized=1]: 11931   <conf:(1)> lift:(2.17) lev:(0.2) conv:(644.84)
3. [native-country= United-States_binarized=1, relationship= Husband_binarized=1]: 11861 ==> [marital-status= Married-civ-spouse_binarized=1]: 11852   <conf:(1)> lift:(2.17) lev:(0.2) conv:(640.57)
4. [native-country= United-States_binarized=1, marital-status= Married-civ-spouse_binarized=1]: 13368 ==> [race= White_binarized=1]: 12369   <conf:(0.93)> lift:(1.08) lev:(0.03) conv:(1.95)
5. [race= White_binarized=1, marital-status= Married-civ-spouse_binarized=1]: 13410 ==> [native-country= United-States_binarized=1]: 12369   <conf:(0.92)> lift:(1.03) lev:(0.01) conv:(1.34)
```

- **FPGrowth 比較與心得:**

執行 FPGrowth 時，明顯發現速度快了不少，FPGrowth 算是個遞迴演算法，期間需要反覆遍歷樹和建構 fp-tree，但是在直接計算單路徑產生所有組合的時候就會便捷很多。

執行結果在 support=0.8 時產生了 2 個關聯法則，和 Apriori 的結果一樣，都是 United Stated->white 和 white->United Stated。

觀察 support 往下調後產生更多的關聯法則，發現結果和 Apriori 幾乎一樣，都是發現(美國人，男人，白人，在私人企業工作)這些特徵關聯特別強。

反思 Apriori 與 FPgrowth 的結果，原本預期以為能發現工作和教育程度的關聯，或是年紀與婚姻之間的關係，結果在 support 較高的條件下幾乎沒有出現。我了解到資料集對於關聯法則有著重要的影響，使得某些潛在的關聯可能會被資料出現機率被過濾掉，如果想要對找出特定關聯的話，可能對資料集的選擇十分重要，亦或是資料集要經過前處理，對於影響演算法裡面權重太多的資料做 sample 或是正規化。

- 利用程式實作和 **Weka** 的結果來探討最後的問題: **(high,low)**與**(support,confidence)**各項組合產生的關聯法則
    1. High support, High confidence:
       美國國籍，白人
    2. High support, low confidence
       白人，美國國籍
    3. Low support, low confidence
       白人，私人企業工作
    4. Low support, high confidence
       丈夫，已經結婚