Machine Learning　HW5

312512005 黃名諄

1. Sequential Forward Selection, SFS :

Index	Feature names
0	mean radius
1	mean texture
2	mean perimeter
3	mean area
4	mean smoothness
5	mean compactness
6	mean concavity
7	mean concave points
8	mean symmetry
9	mean fractal dimension
10	radius error
11	texture error
12	perimeter error
13	area error
14	smoothness error
15	compactness error
16	concavity error
17	concave points error
18	symmetry error
19	fractal dimension error
20	worst radius
21	worst texture
22	worst perimeter
23	worst area
24	worst smoothness
25	worst compactness
26	worst concavity
27	worst concave points
28	worst symmetry
29	worst fractal dimension

上表是各 feature 及其對應 index 的關係，後面的結果中會使用 index 代替 feature name; 後面結果中分類率都取到小數點後第二位

- SFS Highest validated balanced accuracy result of each step :

```
Step1 :
The number of features in this subset:  1
feature subset: (22,)
Highest validated balanced accuracy: 91.21%

Step2 :
The number of features in this subset:  2
feature subset: (22, 24)
Highest validated balanced accuracy: 94.38%

Step3 :
The number of features in this subset:  3
feature subset: (22, 24, 28)
Highest validated balanced accuracy: 95.43%

Step4 :
The number of features in this subset:  4
feature subset: (22, 24, 28, 20)
Highest validated balanced accuracy: 95.78%

Step5 :
The number of features in this subset:  5
feature subset: (22, 24, 28, 20, 0)
Highest validated balanced accuracy: 95.78%

Step6 :
The number of features in this subset:  6
feature subset: (22, 24, 28, 20, 0, 21)
Highest validated balanced accuracy: 96.31%

Step7 :
The number of features in this subset:  7
feature subset: (22, 24, 28, 20, 0, 21, 8)
Highest validated balanced accuracy: 96.31%

Step8 :
The number of features in this subset:  8
feature subset: (22, 24, 28, 20, 0, 21, 8, 2)
Highest validated balanced accuracy: 96.31%

Step9 :
The number of features in this subset:  9
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19)
Highest validated balanced accuracy: 96.49%

Step10 :
The number of features in this subset:  10
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5)
Highest validated balanced accuracy: 96.49%
```

```
Step11 :
The number of features in this subset:  11
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11)
Highest validated balanced accuracy: 96.49%

Step12 :
The number of features in this subset:  12
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4)
Highest validated balanced accuracy: 96.31%

Step13 :
The number of features in this subset:  13
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13)
Highest validated balanced accuracy: 96.13%

Step14 :
The number of features in this subset:  14
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3)
Highest validated balanced accuracy: 96.13%

Step15 :
The number of features in this subset:  15
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27)
Highest validated balanced accuracy: 96.84%

Step16 :
The number of features in this subset:  16
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12)
Highest validated balanced accuracy: 97.01%

Step17 :
The number of features in this subset:  17
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14)
Highest validated balanced accuracy: 97.19%

Step18 :
The number of features in this subset:  18
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6)
Highest validated balanced accuracy: 97.19%

Step19 :
The number of features in this subset:  19
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15)
Highest validated balanced accuracy: 97.36%

Step20 :
The number of features in this subset:  20
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9)
Highest validated balanced accuracy: 97.36%
```

```
Step21 :
The number of features in this subset:  21
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1)
Highest validated balanced accuracy: 97.01%

Step22 :
The number of features in this subset:  22
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18)
Highest validated balanced accuracy: 96.66%

Step23 :
The number of features in this subset:  23
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26)
Highest validated balanced accuracy: 96.49%

Step24 :
The number of features in this subset:  24
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26, 29)
Highest validated balanced accuracy: 96.66%

Step25 :
The number of features in this subset:  25
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26, 29, 25)
Highest validated balanced accuracy: 96.66%

Step26 :
The number of features in this subset:  26
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26, 29, 25, 7)
Highest validated balanced accuracy: 96.31%

Step27 :
The number of features in this subset:  27
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26, 29, 25, 7, 16)
Highest validated balanced accuracy: 96.31%

Step28 :
The number of features in this subset:  28
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26, 29, 25, 7, 16, 17)
Highest validated balanced accuracy: 95.96%

Step29 :
The number of features in this subset:  29
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26, 29, 25, 7, 16, 17, 10)
Highest validated balanced accuracy: 95.61%

Step30 :
The number of features in this subset:  30
feature subset: (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15, 9, 1, 18, 26, 29, 25, 7, 16, 17, 10, 23)
Highest validated balanced accuracy: 95.61%
```

- SFS optimal feature subset :

```
The number of features in the Optimal feature subset: 19
Optimal feature subset:  (22, 24, 28, 20, 0, 21, 8, 2, 19, 5, 11, 4, 13, 3, 27, 12, 14, 6, 15)
Optimal feature subset names:  ['worst perimeter' 'worst smoothness' 'worst symmetry' 'worst radius'
 'mean radius' 'worst texture' 'mean symmetry' 'mean perimeter'
 'fractal dimension error' 'mean compactness' 'texture error'
 'mean smoothness' 'area error' 'mean area' 'worst concave points'
 'perimeter error' 'smoothness error' 'mean concavity' 'compactness error']
Best CR by SFS: 97.36%
```

2. Fisher’s Criterion :

- Top-N-ranked features 之 Validated balanced accuracy :

```
Top-1-ranked features, Validated balanced accuracy: 91.21%
Top-2-ranked features, Validated balanced accuracy: 94.20%
Top-3-ranked features, Validated balanced accuracy: 94.73%
Top-4-ranked features, Validated balanced accuracy: 94.02%
Top-5-ranked features, Validated balanced accuracy: 94.90%
Top-6-ranked features, Validated balanced accuracy: 94.90%
Top-7-ranked features, Validated balanced accuracy: 94.55%
Top-8-ranked features, Validated balanced accuracy: 94.73%
Top-9-ranked features, Validated balanced accuracy: 94.73%
Top-10-ranked features, Validated balanced accuracy: 95.08%
Top-11-ranked features, Validated balanced accuracy: 94.55%
Top-12-ranked features, Validated balanced accuracy: 95.08%
Top-13-ranked features, Validated balanced accuracy: 94.90%
Top-14-ranked features, Validated balanced accuracy: 95.08%
Top-15-ranked features, Validated balanced accuracy: 94.20%
Top-16-ranked features, Validated balanced accuracy: 94.55%
Top-17-ranked features, Validated balanced accuracy: 95.61%
Top-18-ranked features, Validated balanced accuracy: 95.96%
Top-19-ranked features, Validated balanced accuracy: 96.31%
Top-20-ranked features, Validated balanced accuracy: 95.96%
Top-21-ranked features, Validated balanced accuracy: 95.78%
Top-22-ranked features, Validated balanced accuracy: 95.61%
Top-23-ranked features, Validated balanced accuracy: 95.61%
Top-24-ranked features, Validated balanced accuracy: 95.43%
Top-25-ranked features, Validated balanced accuracy: 95.61%
Top-26-ranked features, Validated balanced accuracy: 95.61%
Top-27-ranked features, Validated balanced accuracy: 95.43%
Top-28-ranked features, Validated balanced accuracy: 95.61%
Top-29-ranked features, Validated balanced accuracy: 95.78%
Top-30-ranked features, Validated balanced accuracy: 95.61%
```

- Fisher’s Criterion optimal feature subset :

```
The number of features in the Optimal feature subset: 19
Optimal feature subset:  (27, 22, 7, 20, 2, 23, 0, 3, 6, 26, 5, 25, 10, 12, 13, 21, 24, 28, 1)
Optimal feature subset names:  ['worst concave points' 'worst perimeter' 'mean concave points'
 'worst radius' 'mean perimeter' 'worst area' 'mean radius' 'mean area'
 'mean concavity' 'worst concavity' 'mean compactness' 'worst compactness'
 'radius error' 'perimeter error' 'area error' 'worst texture'
 'worst smoothness' 'worst symmetry' 'mean texture']
Best CR by Fisher's Criterion: 96.31%
```

3. 結果討論 :

1. Sequential Forward Selection 和 Fisher's Criterion 分別屬於 Filter-based 和 Wrapper-based 中的何種特徵篩選方法？

   Ans :

   根據上課所學，Sequential Forward Selection 是 Wrapper-based 的特徵篩選方法，而 Fisher's Criterion 是 Filter-based 的特徵篩選方法

2. 一般來說 Filter-based 和 Wrapper-based 各有什麼性質或優缺點？

   Ans :

   i. Filter-based :

   Filter-based 方法使用預先設立的指標去評估每個 feature 的重要性，像 Fisher 法是計算 F-score，而這樣的方法不需要計算分類率，也因此不會受到選擇不同機器學習算法的影響，並且計算簡單而易於實作，但也有缺點，其只考慮各單一特徵的表現，而沒能考慮到特徵之間互相作用的表現，或許兩個 weak features 一起用卻會有好的表現，這是 Filter-based 考慮不到的。

   ii. Wrapper-based :

   Wrapper-based 方法實際使用分類率篩選，使用表現直接篩選較為直觀，且其有考慮到不同 features 組合間的交互作用表現，但也因此綁定了分類器，使用不同分類器可能帶來不同結果，並且計算上不如 Filter-based 簡單而計算成本較高，另外在過程中被捨棄的 features 沒辦法再被使用評估，還是可能忽略一些潛在的好特徵組合。

3. 在本次作業的結果中是否有展現出跟上一題你的回答有一致的現象呢? 不管是否一致皆請你試著討論與分析原因

   Ans :

在本次作業中能感受到最大差別的是計算效率上，Fisher's Criterion 的計算速度確實快於 SFS，因其需要計算 2-fold 分類率的次數少了很多。另一方面，兩個方法選出的 optimal feature subset 確實不同，表兩方法對特徵篩選的標準上確實不同。我也發現到更改 dataset 的 2-fold 分法，兩個方法的結果都會跟著改變，我認為是因計算分類率上還是會被你的 dataset 影響到，因為兩者都有計算分類率的環節，都會被不同 dataset 拆法影響，如下我改了拆分的 seed，兩者的結果都和之前不同，但可以發現到 SFS 選出的最佳特徵集合分類率都較高，可顯示出 Wrapper-based 如上一題所述考慮特徵間交互作用是很重要的，能選出更佳的特徵組合。

- 更改 2-fold 分法 seed 後的 SFS 結果:

```
The number of features in the Optimal feature subset: 9
Optimal feature subset:  (27, 20, 1, 21, 23, 14, 28, 15, 3)
Optimal feature subset names:  ['worst concave points' 'worst radius' 'mean texture' 'worst texture'
 'worst area' 'smoothness error' 'worst symmetry' 'compactness error'
 'mean area']
Best CR by SFS: 97.54%
```

- 更改 2-fold 分法 seed 後的 Fisher's Criterion 結果:

```
The number of features in the Optimal feature subset: 27
Optimal feature subset:  (27, 22, 7, 20, 2, 23, 0, 3, 6, 26, 5, 25, 10, 12, 13, 21, 24, 28, 1, 17, 4, 8, 29, 15, 16, 19, 14)
Optimal feature subset names:  ['worst concave points' 'worst perimeter' 'mean concave points'
 'worst radius' 'mean perimeter' 'worst area' 'mean radius' 'mean area'
 'mean concavity' 'worst concavity' 'mean compactness' 'worst compactness'
 'radius error' 'perimeter error' 'area error' 'worst texture'
 'worst smoothness' 'worst symmetry' 'mean texture' 'concave points error'
 'mean smoothness' 'mean symmetry' 'worst fractal dimension'
 'compactness error' 'concavity error' 'fractal dimension error'
 'smoothness error']
Best CR by Fisher's Criterion: 95.60%
```