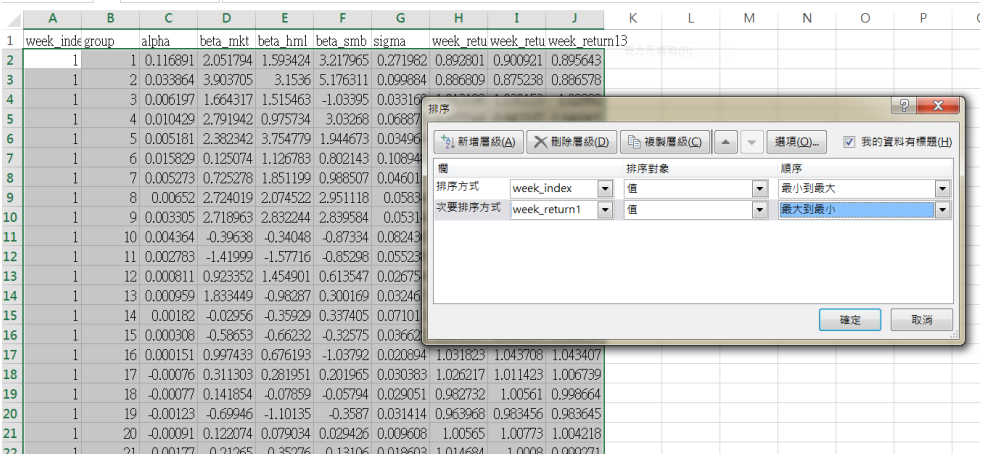
Generative Classification Models

The Homework 3 report from NTU102-1 DMIR course

by NTU Michael Hsu

Data Pre-process

1. 新增欄位 `index` ：原本資料的排序（ = sort by `week_index` and `group` ）。
2. 排序與篩選 `week_index` + `week_return1` ：



	A	B	C	D	E	F	G	H	I	J
1	week_inde	group	alpha	beta_mkt	beta_hml	beta_smb	sigma	week_retu	week_retu	week_return13
2	1	1	0.116891	2.051794	1.593424	3.217965	0.271982	0.892801	0.900921	0.895643
3	1	2	0.033864	3.903705	3.1536	5.176311	0.099884	0.886809	0.875238	0.886578
4	1	3	0.006197	1.664317	1.515463	-1.03395	0.03316			
5	1	4	0.010429	2.791942	0.975734	3.03268	0.06887			
6	1	5	0.005181	2.382342	3.754779	1.944673	0.03496			
7	1	6	0.015829	0.125074	1.126783	0.802143	0.10894			
8	1	7	0.005273	0.725278	1.851199	0.988507	0.04601			
9	1	8	0.00652	2.724019	2.074522	2.951118	0.0583			
10	1	9	0.003305	2.718963	2.832244	2.839584	0.0531			
11	1	10	0.004364	-0.39638	-0.34048	-0.87334	0.08243			
12	1	11	0.002783	-1.41999	-1.57716	-0.85298	0.05523			
13	1	12	0.000811	0.923352	1.454901	0.613547	0.02675			
14	1	13	0.000959	1.833449	-0.98287	0.300169	0.03246			
15	1	14	0.00182	-0.02956	-0.35929	0.337405	0.07101			
16	1	15	0.000308	-0.58653	-0.66232	-0.32575	0.03662			
17	1	16	0.000151	0.997433	0.676193	-1.03792	0.020894	1.031823	1.043708	1.043407
18	1	17	-0.00076	0.311303	0.281951	0.201965	0.030383	1.026217	1.011423	1.006739
19	1	18	-0.00077	0.141854	-0.07859	-0.05794	0.029051	0.982732	1.00561	0.998664
20	1	19	-0.00123	-0.69946	-1.10135	-0.3587	0.031414	0.963968	0.983456	0.983645
21	1	20	-0.00091	0.122074	0.079034	0.029426	0.009608	1.00565	1.00773	1.004218

3. 定義分類標簽：
    - 1. 新增欄位 `index_sort` ：根據上一個步驟後的排序。
    - 1. 新增欄位 `index_sort % 30` ： `mod(左邊, 30)`
    - 1. 給予分類標籤 `class` ： `=IF((左邊>0)*(左邊<=6),"1","0")` 前六個為 1，剩下二十四個為 0。
4. 新增欄位 `random_sort` ：最後依據這個欄位 `=RAND()` 來做 10-fold classification。
5. 最後整理資料為 `data/ldpa30_train use.csv`
    - 剩下 feature

`alpha` 、 `beta_mkt` 、 `beta_hml` 、 `beta_smb` 、 `sigma`

- 分類的標簽 `class`
- 以及目前的隨機排序依據，作為切割十份用，產生新的 `new_index` 。

new_index	alpha	beta_mkt	beta_hml	beta_smb	sigma	class
1	-0.0065066	3.048392	4.07517039	0.9617558	0.01808332	0
2	0.00057031	0.32789225	0.4956953	0.59191614	0.06036168	0
3	0.00236977	1.44659396	-1.835827	2.10289862	0.01848847	0
4	0.00019669	1.51353161	1.08480384	1.94350839	0.01727014	0
5	-0.0042819	1.45120471	1.86734677	0.59260946	0.02066149	0

如何執行

R cmd:

```
> source("/path_to/generative_classification_model.r")
```

example: （可用拖曳方式取得路徑）

```
> source("/Users/michaelhsu/Dropbox/15.\ 碩一上課業/02.\ DMIR\
```

結果 (10-fold-validation)

```
        accuracy precision     recall  F-measure
bin1   0.7990991 0.2857143 0.01843318 0.03463203
bin2   0.7918919 0.4090909 0.03964758 0.07228916
bin3   0.7972973 0.4444444 0.01785714 0.03433476
bin4   0.8099099 0.5454545 0.02830189 0.05381166
bin5   0.7873874 0.3571429 0.02155172 0.04065041
bin6   0.8045045 0.2500000 0.01913876 0.03555556
bin7   0.8009009 0.5000000 0.04072398 0.07531381
bin8   0.8045045 0.5000000 0.03686636 0.06866953
bin9   0.7954955 0.5000000 0.04405286 0.08097166
bin10  0.7945946 0.7142857 0.04273504 0.08064516
         accuracy precision     recall  F-measure
mean  0.798558559 0.4506133 0.03093085 0.05768737
sd    0.006677209 0.1348956 0.01095288 0.01996029
```

Source code

https://github.com/evenchange4/102-1_DMIR_Hw3_Generative-Classification-Models

Reference

- row bind example
- subset a data set meet a condition