期中報告

科技業半導體個股股價預測

G11
林語萱、何亞凡、吳欣婷
廖韋凱、陳廷旭、游羽琛

Agenda

1. 關鍵字向量空間建立
2. 機器學習
3. 移動回測
4. 結語

Agenda

1. 關鍵字向量空間建立
2. 機器學習
3. 移動回測
4. 結語

個股選擇

篩選準則：討論量大、股價波動大

台積電

聯電

聯發科

大盤指數

種子關鍵字設定

設定標準：人工觀察文章列表中的漲跌相關字詞

看漲關鍵字

上漲、漲停、漲幅、利多、成長、
加碼、上升、漲勢、上看、紅盤、
收紅、買超、攀升、高點、高峰、
新高、降息、多方、接單、上車、
獲利、樂觀、護盤、布局、漲、賺

看跌關鍵字

下跌、下挫、跌停、跌幅、利空、
衰減、衰退、減碼、跌勢、綠盤、
低點、收綠、收黑、重挫、減產、
賣超、再跌、新低、摜破、空方、
停工、悲觀、斷頭、損失、跌、賠、
虧、黑天鵝、信用破產

建構看漲/看跌文章集

以聯電為例

- 依據標題與內文過濾出包含**聯電**或**洪嘉聰**之文章

- 對於每一文章，若**標題**出現種子關鍵字詞：
  - 至少含有一個看漲詞，沒有任何看跌詞：把文章歸類在**看漲文章集**
  - 至少含有一個看跌詞，沒有任何看漲詞：把文章歸類在**看跌文章集**
  - 同時包含看漲詞和看跌詞：捨棄該文章
  - 沒有任何看漲詞或看跌詞：捨棄該文章

從文章集篩選出關鍵字

- 讀入 stopword list
  - stopword list 來源https://github.com/tomlinNTUB/Python-in-5-days/blob/master/10-2%20%E4%B8%AD%E6%96%87%E6%96%B7%E8%A9%9E-%E7%A7%BB%E9%99%A4%E5%81%9C%E7%94%A8%E8%A9%9E.md

- 分別Preprocess 看漲/看跌文章集的每篇文章
  - 去除英文、數字與含特殊符號
  - 去除停用詞
  - 累計每個 gram 的 TF 及 DF
  - 去除 DF <= 10 的 gram

從文章集篩選出關鍵字

- 各主題相關文章取出 2 ~ 6 gram，計算 TF、DF
- 語料庫中所有文章，取出 2 ~ 6 gram，計算 TF、DF
- 對每個留下的 gram，合併可能多餘的子字串，得出最終 gram list
  - 合併標準：
    若 |DF(n gram)-DF(n+1 gram)| / DF(n gram) < 0.02
    則移除該n gram (n = 2 ~ 5)

- 計算 gram list 中每個 gram 之 TF-IDF, all_TF_IDF, expected_TF, expected_DF, DF_chisquare, MI, Lift

關鍵字範例：聯電看漲、看跌

gram	TF_IDF
驅動	4.4794507
法院	4.3574035
再生	4.3543652
日圓	4.3251878
歐盟	4.3215705
護盤	4.2692968
晶圓代工產	4.2417423
庫藏股	4.2416597
敦泰	4.2166329
電源	4.2023545

聯電看漲關鍵字（部分）

gram	TF_IDF
外交	5.203134658
下跌點或收點	5.067475117
股賣	5.018469261
收購	5.006145838
金價	4.984157774
賣超較多	4.970741346
認列	4.958631394
日圓	4.952243747
處分	4.926898367
夏普	4.926898367

聯電看跌關鍵字（部分）

Agenda

1 關鍵字向量空間建立

2 機器學習

3 移動回測

4 結語

Step1：調整參數

- 看漲/跌文章集中各自嘗試選取不同關鍵字數目
  - top k [300,500,700]

- 嘗試不同指標篩選出的關鍵字
  - top k index ['TF_IDF','TF','TF*DF_chisquare','MI','Lift']

- 嘗試有無刪除看漲看跌重複關鍵字
  - remove in [False, True]

Step2：調整資料、以8：2分配進行訓練測試

- Resampling
  處理不平衡資料
  將看漲與看跌文章數目調整至一致

- Normalization
  把資料分布做正規化，減少不同特徵變異的影響

- Train: test: 8:2
  以80%訓練資料、20%測試資料的比例，進行機器學習

Step3：機器學習模型選擇

- 四間公司分別各嘗試三種機器學習模型
  - Random Forest (n_estimators = [500, 1000, 2000])
    以500, 1000, 2000三種不同樹數目測試
  - K-Nearest Neighbor (n_neighbors = [5,9,13])
    以5, 9, 13三種不同鄰居數測試
  - Support Vector Machine

各模型測試最佳結果 - 聯電

86%

RF	pred. 漲	pred. 跌
Actual 漲	190	23
Actual 跌	28	127

top k：300
index：TF_IDF
remove：False
n_estimator：1000

83%

SVM	pred. 漲	pred. 跌
Actual 漲	179	34
Actual 跌	28	127

top k：500
index：TF_IDF
remove：False
n_estimator：N/A

80%

KNN	pred. 漲	pred. 跌
Actual 漲	179	34
Actual 跌	41	144

top k：300
index：TF_IDF
remove：True
n_neighbor：5

各模型測試最佳結果 - 台積電

84%

RF	pred. 漲	pred. 跌
Actual 漲	546	114
Actual 跌	116	655

top k：300
index：TF_IDF
remove：False
n_estimator：1000

84%

SVM	pred. 漲	pred. 跌
Actual 漲	536	124
Actual 跌	105	666

top k：400
index：TF_IDF
remove：False
n_estimator：N/A

70%

KNN	pred. 漲	pred. 跌
Actual 漲	514	146
Actual 跌	286	485

top k：300
index：TF_IDF
remove：True
n_neighbor：5

各模型測試最佳結果 - 聯發科

83%

RF	pred. 漲	pred. 跌
Actual 漲	140	33
Actual 跌	28	162

top k：700
index：TF_IDF
remove：False
n_estimator：1000

82%

SVM	pred. 漲	pred. 跌
Actual 漲	135	38
Actual 跌	29	161

top k：700
index：TF_IDF
remove：False
n_estimator：N/A

73%

KNN	pred. 漲	pred. 跌
Actual 漲	148	25
Actual 跌	73	117

top k：700
index：TF_IDF
remove：False
n_estimator：5

各模型測試最佳結果 - 大盤指數

87%

RF	pred. 漲	pred. 跌
Actual 漲	582	118
Actual 跌	83	816

top k：500
index：TF_IDF
remove：False
n_estimator：2000

86%

SVM	pred. 漲	pred. 跌
Actual 漲	596	104
Actual 跌	120	779

top k：500
index：TF_IDF
remove：False
n_estimator：N/A

72%

KNN	pred. 漲	pred. 跌
Actual 漲	552	148
Actual 跌	305	594

top k：500
index：TF_IDF
remove：False
n_neighbor：5

機器學習模型比較

- 達成較佳 F1值 (83%~87%) 的條件：
  - **模型：RF (隨機森林)**
  - **指標：TF-IDF**
  - **不刪除看漲看跌重複關鍵字**

- 其它如模型參數的選擇、字詞使用數量，隨著不同股票而有差異

Agenda

1 關鍵字向量空間建立

2 機器學習

3 移動回測

4 結語

移動回測方法

- 在36個月的資料當中
  過濾出欲預測月份**前3或5個月**的文章集
- 分別試用RF、SVM建構訓練模型
- 預測該月**每一天**的股價漲跌

移動回測結果 - 台積電（RF/SVM）

台積電 3-month			
	Real rise	Real down	total
RF rise	84	119	203
RF down	213	196	409
RF 不出手	34	33	67
total	331	348	679
出手	90.13%		
準確	45.75%		

台積電 3-month			
	Real rise	Real down	total
SVM rise	89	111	200
SVM down	203	199	402
SVM 不出手	39	38	77
total	331	348	679
出手	88.66%		
準確	47.84%		

台積電 5-month			
	Real rise	Real down	total
RF rise	54	81	135
RF down	226	216	442
RF 不出手	33	29	62
total	313	326	639
出手	90.30%		
準確	46.79%		

台積電 5-month			
	Real rise	Real down	total
SVM rise	89	108	197
SVM down	197	184	381
SVM 不出手	27	34	61
total	313	326	639
出手	90.45%		
準確	47.23%		

移動回測結果 - 聯電（RF/SVM）

聯電 3-month			
	Real rise	Real down	total
RF rise	48	65	113
RF down	124	166	290
RF 不出手	107	169	276
total	279	400	679
出手	59.35%		
準確	53.10%		

聯電 3-month			
	Real rise	Real down	total
SVM rise	36	54	90
SVM down	145	208	353
SVM 不出手	98	138	236
total	279	400	679
出手	65.24%		
準確	55.08%		

聯電 5-month			
	Real rise	Real down	total
RF rise	46	54	100
RF down	120	164	284
RF 不出手	99	156	255
total	265	374	639
出手	60.09%		
準確	54.69%		

聯電 5-month			
	Real rise	Real down	total
SVM rise	31	44	75
SVM down	140	198	338
SVM 不出手	94	132	226
total	265	374	639
出手	64.63%		
準確	55.45%		

移動回測結果 - 聯發科（RF/SVM）

聯發科 3-month			
	Real rise	Real down	total
RF rise	96	111	207
RF down	135	146	281
RF 不出手	84	107	191
total	315	364	679
出手	71.87%		
準確	49.59%		

聯發科 3-month			
	Real rise	Real down	total
SVM rise	58	64	122
SVM down	178	186	364
SVM 不出手	79	114	193
total	315	364	679
出手	71.58%		
準確	50.21%		

聯發科 5-month			
	Real rise	Real down	total
RF rise	90	105	195
RF down	127	140	267
RF 不出手	76	101	177
total	293	346	639
出手	72.30%		
準確	49.78%		

聯發科 5-month			
	Real rise	Real down	total
SVM rise	36	44	80
SVM down	188	201	389
SVM 不出手	69	101	170
total	293	346	639
出手	73.40%		
準確	50.53%		

移動回測結果 - 大盤（RF/SVM）

大盤 3-month			
	Real rise	Real down	
RF rise	123	122	245
RF down	211	151	362
RF 不出手	37	35	72
total	371	308	679
出手	89.40%		
準確	45.14%		

大盤 3-month			
	Real rise	Real down	total
SVM rise	141	120	261
SVM down	184	152	336
SVM 不出手	46	36	82
total	371	308	679
出手	87.92%		
準確	49.08%		

大盤 5-month			
	Real rise	Real down	total
RF rise	104	95	199
RF down	216	157	373
RF 不出手	33	34	67
total	353	286	639
出手	89.51%		
準確	45.63%		

大盤 5-month			
	Real rise	Real down	total
SVM rise	135	112	247
SVM down	180	140	320
SVM 不出手	38	34	72
total	353	286	639
出手	88.73%		
準確	48.50%		

移動回測 Another Try

- **重新建構資料集**：
  先篩出各公司每月股價「漲幅最大」和「跌幅最大」的 top 5天
  以這些天的「前一天」的文章作為公司文章集
  以此過濾出欲預測月份**前3或5個月**的文章集

- 分別試用RF、SVM建構訓練模型

- 預測該月**每一天**的股價漲跌

Another Try 結果 - 台積電（RF/SVM）

台積電 3-month			
	Real rise	Real down	total
RF rise	144	159	303
RF down	160	172	332
RF 不出手	11	33	44
total	315	364	679
出手	93.52%		
準確	49.76%		

台積電 3-month			
	Real rise	Real down	total
SVM rise	137	167	304
SVM down	153	170	323
SVM 不出手	25	27	52
total	315	364	679
出手	92.34%		
準確	48.96%		

台積電 5-month			
	Real rise	total	total
RF rise	155	162	317
RF down	130	157	287
RF 不出手	30	43	73
total	315	362	677
出手	89.22%		
準確	51.66%		

台積電 5-month			
	Real rise	Real down	total
SVM rise	160	180	340
SVM down	124	136	260
SVM 不出手	31	46	77
total	315	362	677
出手	88.63%		
準確	49.33%		

移動加權平均法、技術分析

移動加權平均法		
Predict	Accuracy	Shot rate
n=2	53.06%	100%
n=3	52.32%	100%
n=4	53.62%	100%

技術分析		
Interval	Accuracy	Shot rate
100	58.80%	27.70%
30	87.50%	1.11%

移動回測方法比較

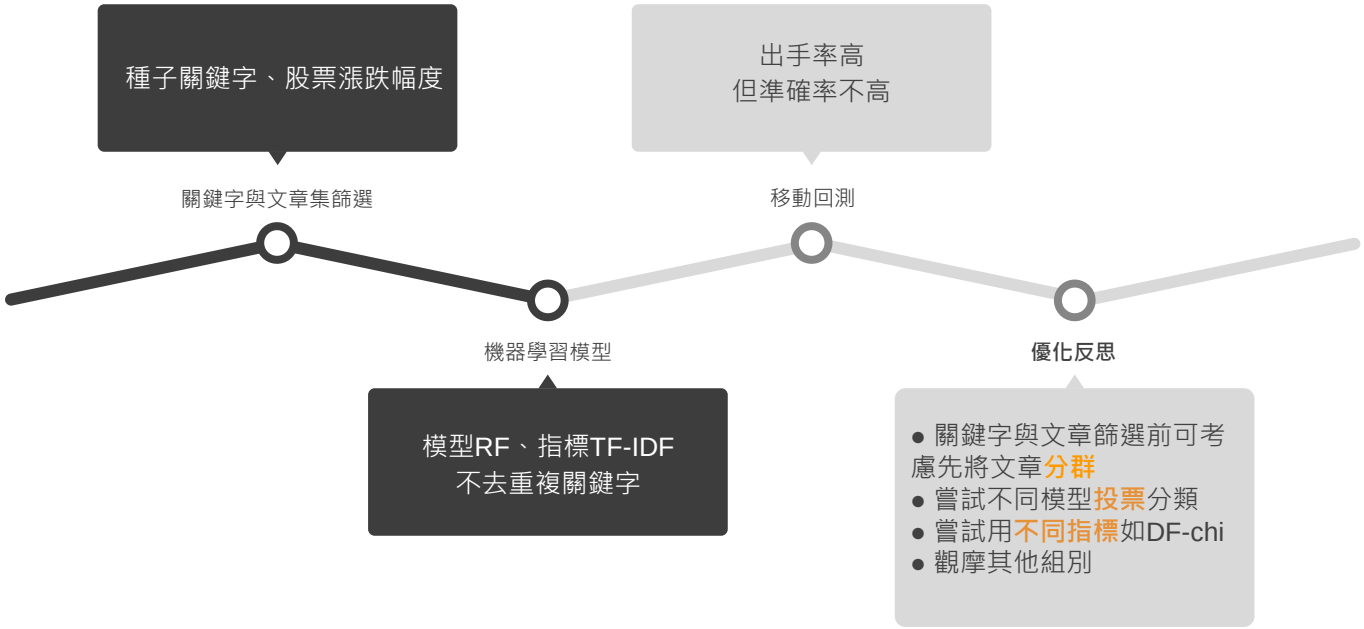
- 兩種方法得到的文章資料集：
  - 種子關鍵字篩選：出手率**高**、準確率**較低**
  - 股票漲跌幅度篩選：出手率、準確率皆**略為提高**
  - 兩種方法的準確率都只在50%附近
  - 在我們的模型下，預測股價為跌的機會比較高

- 移動加權平均法：出手率可達100%、準確率**仍低**
- 技術分析：準確率**可以提高許多**，但需**犧牲出手率**

Agenda

1. 關鍵字向量空間建立
2. 機器學習
3. 移動回測
4. 結語

總結：專案進行流程


種子關鍵字、股票漲跌幅度
關鍵字與文章集篩選
機器學習模型
模型RF、指標TF-IDF
不去重複關鍵字
出手率高
但準確率不高
移動回測
優化反思
● 關鍵字與文章篩選前可考慮先將文章分群
● 嘗試不同模型投票分類
● 嘗試用不同指標如DF-chi
● 觀摩其他組別

附錄：機器模型F1值數值計算 - 聯電

Random Forest

	precision	recall	f1-score	support
0.0	0.87	0.89	0.88	213
1.0	0.85	0.82	0.83	155
accuracy			0.86	368
macro avg	0.86	0.86	0.86	368
weighted avg	0.86	0.86	0.86	368

Support Vector Machine

	precision	recall	f1-score	support
0.0	0.86	0.84	0.85	213
1.0	0.79	0.82	0.80	155
accuracy			0.83	368
macro avg	0.83	0.83	0.83	368
weighted avg	0.83	0.83	0.83	368

K-Nearest-Neighbor

	precision	recall	f1-score	support
0.0	0.81	0.84	0.83	213
1.0	0.77	0.74	0.75	155
accuracy			0.80	368
macro avg	0.79	0.79	0.79	368
weighted avg	0.80	0.80	0.80	368

附錄：機器模型F1值數值計算 - 台積電

Random Forest

	precision	recall	f1-score	support
0.0	0.82	0.83	0.83	660
1.0	0.85	0.85	0.85	771
accuracy			0.84	1431
macro avg	0.84	0.84	0.84	1431
weighted avg	0.84	0.84	0.84	1431

Support Vector Machine

	precision	recall	f1-score	support
0.0	0.84	0.81	0.82	660
1.0	0.84	0.86	0.85	771
accuracy			0.84	1431
macro avg	0.84	0.84	0.84	1431
weighted avg	0.84	0.84	0.84	1431

K-Nearest-Neighbor

	precision	recall	f1-score	support
0.0	0.64	0.78	0.70	660
1.0	0.77	0.63	0.69	771
accuracy			0.70	1431
macro avg	0.71	0.70	0.70	1431
weighted avg	0.71	0.70	0.70	1431

附錄：機器模型F1值數值計算 - 聯發科

Random Forest

	precision	recall	f1-score	support
0.0	0.83	0.80	0.81	173
1.0	0.82	0.85	0.84	190
accuracy			0.83	363
macro avg	0.83	0.83	0.83	363
weighted avg	0.83	0.83	0.83	363

Support Vector Machine

	precision	recall	f1-score	support
0.0	0.82	0.78	0.80	173
1.0	0.81	0.85	0.83	190
accuracy			0.82	363
macro avg	0.82	0.81	0.81	363
weighted avg	0.82	0.82	0.82	363

K-Nearest-Neighbor

	precision	recall	f1-score	support
0.0	0.67	0.86	0.75	173
1.0	0.82	0.62	0.70	190
accuracy			0.73	363
macro avg	0.75	0.74	0.73	363
weighted avg	0.75	0.73	0.73	363

附錄：機器模型F1值數值計算 - 大盤

Random Forest

	precision	recall	f1-score	support
0.0	0.88	0.83	0.85	700
1.0	0.87	0.91	0.89	899
accuracy			0.87	1599
macro avg	0.87	0.87	0.87	1599
weighted avg	0.87	0.87	0.87	1599

Support Vector Machine

	precision	recall	f1-score	support
0.0	0.83	0.85	0.84	700
1.0	0.88	0.87	0.87	899
accuracy			0.86	1599
macro avg	0.86	0.86	0.86	1599
weighted avg	0.86	0.86	0.86	1599

K-Nearest-Neighbor

	precision	recall	f1-score	support
0.0	0.64	0.79	0.71	700
1.0	0.80	0.66	0.72	899
accuracy			0.72	1599
macro avg	0.72	0.72	0.72	1599
weighted avg	0.73	0.72	0.72	1599

附錄：說明影片連結