詹雅筑 B06106017 郭太元 R09725015 陳又加 B06303096 蔡淳如 B06303116 黃筠芝 B06303106 陳泳寧 B06607047

From Classification to Prediction

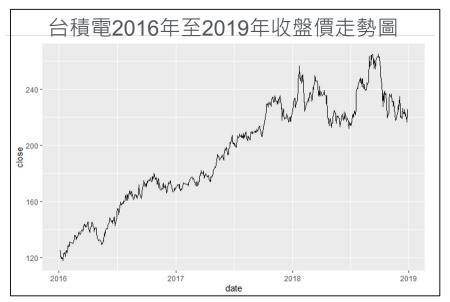
目錄

- 1 選股
- 2 實驗參數設計
- 3 資料前處理
- 建構向量空間

- 5 分類模型
- 6 分類結果
- 7 資料回測
- 8 進階應用

選股





股價波動明顯

整體上漲趨勢

討論熱度高,新聞數量多

未來股價漲跌說法不一,希望能有參考依據

實驗參數設計

固定參數

比較第D+1 天與第 D 天的股價收盤價

若上漲超過 0.1% · 則視第 D 天的文章為一批看漲文件集

若下跌超過 0.1% ,則視第 D 天的文章為一批看跌文件集

$$(n = 1; \sigma = 0.001)$$

資料前處理

- 1. 去除資訊含量低的PTT的公告文
- 針對個股篩選出相關的新聞 只選擇該公司名稱作為關鍵詞的原因在於,本組認為若該公司所發生的事件足以影響該公司股價表現,則報導該事件的新聞必定會提到該公司名稱
- 1. 利用前述固定參數,將文章分類為看漲文章與看跌文章

個股名稱	關鍵字	新聞篇數	浮動	參數
			看漲篇數	看跌篇數
台積電	台積電、臺積電	21419	9964	8095

建構向量空間

- 1. 利用Jeiba斷詞,去除標點符號、英文、數字、停止字詞,留下2~6gram斷詞
- 2. 利用DF_全部IDF作為關鍵字排序指標

分別找出上漲與下跌新聞中前100個個股關鍵字

	關鍵字	
上漲	上市日、拍板、中秋節、新元、驍龍	
下跌	條約、休兵、電腦病毒、感染、熊本	

建構向量空間

3. 建構 Document Term Matrix

新聞ID	分類(pos/neg)	A詞	B詞	
1	pos = 1	出現次數	出現次數	出現次數
2	neg = -1	出現次數	出現次數	出現次數

實際結果範例					
新聞ID	分類(pos/neg)	上市日	新元	電腦病毒	感染
207791	1	1	3	0	0
192693	-1	0	0	5	4

分類模型

Step 1: 資料分類

將看漲及看跌兩批文章分類為訓練資料及測試資料

Training: 80%, Testing: 20%

Step 2: 資料訓練

以6個分類模型訓練資料並進行分類

Logistic Regression / Multinomial Regression / Naive Bayes / SVM / Decision Tree / Random Forest

分類結果

Logistic Regression	真實為漲	真實為跌
預測為漲	401	231
預測為跌	321	704

準確率:66.6868%

SVM	真實為漲	真實為跌
預測為漲	452	255
預測為跌	270	680

準確率:68.3162%

Multinomial Regression	真實為漲	真實為跌
預測為漲	415	244
預測為跌	307	691

準確率:66.7471%

Decision Tree	真實為漲	真實為跌
預測為漲	9	2
預測為跌	713	933

準確率:56.8497%

Naive Bayes	真實為漲	真實為跌
預測為漲	349	226
預測為跌	373	709

準確率:63.8503%

Random Forest	真實為漲	真實為跌
預測為漲	424	279
預測為跌	298	656

準確率:65.1780%

進階應用:股市辭典

將股市辭典中的關鍵字分別加入上漲及下跌關鍵字

	股市辭典		
上漲	成長、獲利、增加、上漲、看好		
下跌	下滑、下跌、減少、衝撃、虧損		

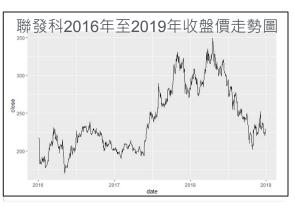
SVM	真實為漲	真實為跌
預測為漲	20	2
預測為跌	635	769

準確率:55.3296%

進 階 應 用: 聯發科

個股名稱	關鍵字	新聞篇數	固定參數	
1 10 次		机耳扁数	看漲篇數	看跌篇數
聯發科	聯發科	6141	2912	2577

	閣 鍵 字		
上漲	加薪、新能源、新元、超出、暴增		
下跌	張陳浩、楠梓、解散、罕見、安泰		



SVM	真實為漲	真實為跌
預測為漲	208	107
預測為跌	103	225

準確率:67.3406%

進階應用:浮動參數

利用浮動參數,將文章分類為看漲文章與看跌文章

不同個股的股票漲幅大不相同,為了能讓個股更精確地與自身股價相比,因此本組設計浮動參數,作為另一種參考值

/D 00 20 150	BB ∆7± ₽	 新聞篇數	固定參數	
個股名稱	個股名稱 關鍵字 關鍵字		看漲篇數	看跌篇數
台積電	台積電、臺積電	21,419	2,186	2,400

	關 鍵 字
上漲	無限期、大晶圓、格芯、就職演說、開票
下跌	電腦病毒、感染、嚴陳莉蓮、病毒感染、熊本

SVM	真實為漲	真實為跌
預測為漲	197	78
預測為跌	61	151

準確率:71.4579%

雖然使用浮動參數提高了SVM模型分類結果的準確率,然而若用浮動參數跑資料回測結果非常不理想,因此這項指標仍有討論空間

資料回測

以SVM模型驗證準確率&出手率

看漲

當日看漲新聞數 - 當日看跌新聞數 當日看漲新聞數 + 當日看跌新聞數 > 0.01

持平

當日看漲新聞數 - 當日看跌新聞數 | ≤ 0.01

看跌

當日看漲新聞數 - 當日看跌新聞數 < -0.01 當日看漲新聞數 + 當日看跌新聞數

SVM	真實為漲	真實為跌
預測為漲	157	12
預測為跌	73	126

出手率

76.2887%

1 - 持平率

準確率

83.5907%

預測漲跌日期與實際漲跌日期相同的比率

進階應用:資料回測

權重法:將每日新聞態度皆納入考量,並以距今幾日作為權重

$$\sum_{n=0}^{\infty}$$

(距今第n日看漲新聞)× 0.2^n - (距今第n日看跌新聞)× 0.2^n (距今第n日看漲新聞)× 0.2^n + (距今第n日看跌新聞)× 0.2^n

$$n = 0$$
 時為今日

出手率

89.5435%

準確率

80.2218%

SVM	真實為漲	真實為跌
預測為漲	157	12
預測為跌	73	126

解說影片連結

https://youtu.be/37SwuTJBeTw

THANKS!