Financial index prediction based on machine learning

基於機器學習進行金融指數預測

組別：B77　組員：劉安得　指導教授：翁詠祿

Introduction

近年來「智能投資」，也就是所謂「機器人理財」逐漸流行。金融監督管理委員會(金管會)也在2017年開放「自動化投資理財顧問」(Robo-advisor)業務。「機器人理財」為利用人工智慧演算法，提供投資者理財建議，甚至進行自動化投資組合調整。本研究嘗試以機器學習進行金融指數的預測，希望未來可以應用在機器人理財上。

System design

1. 資料集準備

(1) 資料下載

本研究使用一個叫做yfinance的函式庫來下載Yahoo! Finance的資料。下載下來的資料會有開盤價、最高價、最低價、收盤價、調整後收盤價、交易量這六筆資料。本研究抓取了Dow Jones工業平均指數、S&P 500、NASDAQ綜合指數、Apple、Tesla、台股加權指數、元大台灣50、台泥、台積電等股票或指數來進行分析。時間則是取2015年1月1日至2020年12月4日。

(2) 內插法

由於同時下載台灣股票與美國股票的資料，但兩國開盤日不盡相同，因此會造成某些資料有缺漏。舉例來說，如果美股有開盤但台股沒開盤的話，那此日台股的資料就會有缺漏。可以使用內插法把缺漏的值補上，本研究使用SciPy函式庫的interpolate函式來進行內插，讓每一筆資料都有個數值。

(3) 技術分析

由於我最後想驗證機器學習是否需要輸入技術分析指標進去，理論上是不需要，因為技術分析指標皆是由股價以及交易量所計算出來，而機器學習模型有辦法自己分析出這些指標。

為了驗證以上所述，因此要先把各技術分析指標計算出來。本研究使用現成的函式庫，計算出RSI5, OBV, 以及各移動平均線(MA)，以供未來使用。

資料預處理

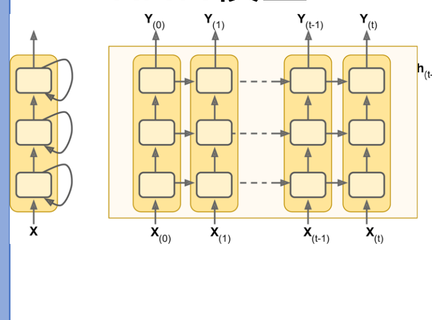
2. 資料預處理

首先先標準化。因為本研究希望透過前五天的資料來預測下一天的股價，然而我們拿到的資料為一整條時間序列，因此要先經過預處理。將一條時間序列五天五天分為一段，因此會得到很多段時間序列，這些就是之後要進行機器學習的資料。最後由於各小段時間序列彼此之間已經不再有順序性了，只有各段那五天資料的順序不能被打亂，因此在分割訓練集與測試集時，可以進行打亂的動作。

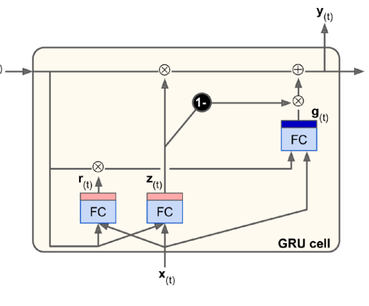
3. Model

一共有三種模型分別進行預測，RNN、LSTM、GRU，每種模型都設置三層layer，最後一層取時間序列最後一筆資料當成預測值。loss函數則選擇為MSE，optimizer則是Adam，一共跑20個epochs。

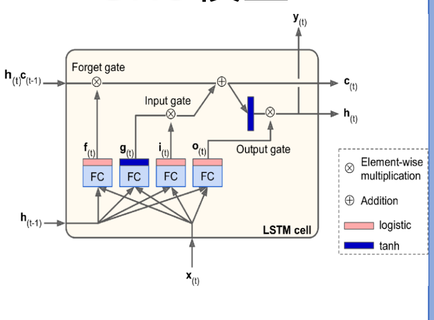
RNN模型



LSTM模型



GRU模型



Result

1. 各模型比較

每一筆資料跑25次，並計算MSE的平均值與標準差

股票代碼	^DJI	^GSPC	^IXIC
RNN	0.00116 +/- 0.00028	0.00093 +/- 0.00027	0.00064 +/- 0.00021
LSTM	0.00131 +/- 0.00021	0.00098 +/- 0.00016	0.00061 +/- 0.00014
GRU	0.00108 +/- 0.00020	0.00078 +/- 0.00020	0.00047 +/- 0.00010
Linear Reg.	0.00207 +/- 0.00149	0.00201 +/- 0.00086	0.00115 +/- 0.00071
股票代碼	**AAPL**	**TSLA**	**^TWII**
RNN	0.00040 +/- 0.00013	0.00033 +/- 0.00011	0.00071 +/- 0.00023
LSTM	0.00032 +/- 0.00007	0.00047 +/- 0.00015	0.00090 +/- 0.00011
GRU	0.00029 +/- 0.00005	0.00031 +/- 0.00012	0.00071 +/- 0.00019
Linear Reg.	0.00113 +/- 0.00050	0.00063 +/- 0.00040	0.00174 +/- 0.00096
股票代碼	**0050.TW**	**1101.TW**	**2330.TW**
RNN	0.00071 +/- 0.00025	0.00100 +/- 0.00043	0.00057 +/- 0.00020
LSTM	0.00069 +/- 0.00013	0.00077 +/- 0.00015	0.00043 +/- 0.00006
GRU	0.00046 +/- 0.00011	0.00063 +/- 0.00017	0.00035 +/- 0.00007
Linear Reg.	0.00209 +/- 0.00117	0.00437 +/- 0.00365	0.00165 +/- 0.00124

2. 技術分析之驗證

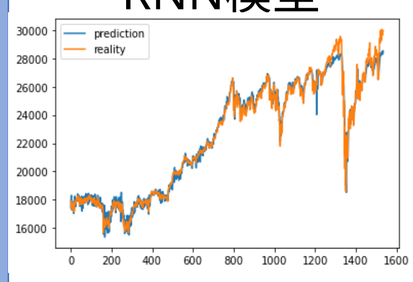
由下表可見，有無技術分析在各模型上並無顯著差異。因此可以推斷機器學習模型不需要輸入技術分析指標，它能自己分析出來。

	無技術分析	有技術分析
RNN	0.00116 +/- 0.00028	0.00124 +/- 0.00053
LSTM	0.00131 +/- 0.00021	0.00109 +/- 0.00019
GRU	0.00108 +/- 0.00020	0.00101 +/- 0.00021
Linear Reg.	0.00207 +/- 0.00149	0.00262 +/- 0.00131

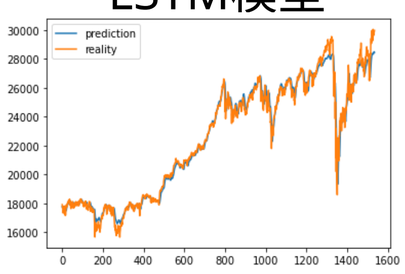
3. 研究結果

皆為預測道瓊指數

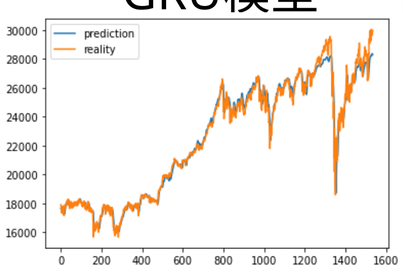
RNN模型



LSTM模型



GRU模型



Conclusion

雖然各股票所預測出來的MSE都非常低，然而光是僅僅猜測前一天收盤價為預測值，MSE就能達到0.000369，跑機器學習模型還不一定表現得比它好。除此之外，就算能精準地預測股價，也不知道如何轉換成交易策略，因此此主題還有許多發展性。本研究現階段為預測下一天股價，未來可以往預測更多天進行。此外，現在也有將強化學習應用在股票交易的論文，未來或許能以強化學習訓練出一套交易策略。

Reference

[1] al., K. C. (2014). Learning Phrase Representations Using RNN Encoder-Decoder for Statistical. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (pp. 1724-1734).