Soft Feature LSTM = SoftfeatLSTM

======================================================================

科技部計畫內容 v.s. 我們真實實行的內容

本次研究著重的重點在於比較 LSTM、基因演算法、LSTM 與基因演算法之組合模型，我們希望能透過兩者的結合找出「最大獲利」、「最高勝率」、以及「最低最大虧損」。以LSTM 來結合圖形辨識做預測，因此在一開始搜集特徵來學習，我們將會運用 CNN 的概念，用特徵提取的方法找出需要的特徵值。

1. 比較 a. LSTM、b. 基因演算法、c. LSTM 與基因演算法 → 需要討論，目前看起來有東西可以呈現

2. 找出「最大獲利」、「最高勝率」、以及「最低最大虧損」 → 我們現在只有看投資報酬率

3. LSTM + Image → 這部分我們沒有實作

======================================================================

時程規劃

（ㄧ）建立海報內容

（二）優化程式碼並加上註解

（三）針對使用者選股模型進行優化去overfit

ㄧ、海報內容與論文架構

A. 摘要：

1. 別人的作法有什麼缺點？技術指標...
2. 我們怎麼優化？

* 四價格一量搭配標準化即佔有技術面的主要特徵
* 為何選擇融資融券跟法人？當有足夠incentive讓你融資融券代表有一定程度的漲跌

B. 資料處理：

1. 資料搜集：先確認研究範圍為0050成分股，但以過往survey的paper來說，僅僅是搜集開盤收盤與最高最低價格是遠遠不夠的，所以我們額外使用融資融券與股市三大法人的買多賣空資訊來整合已進行後續資料處理，但因為加入了融資融券，所以我們資料限縮到了2012開始（自2012 證券交易所才揭露此訊息）。

2. 做exponential smoothing 讓價格及所有特徵平滑更好訓練

* 1. 放圖表說明價格震盪（兩條線：有平滑即沒平滑）
  2. 提供數據說明只有平滑價格以及平滑所有特徵的投資報酬率

3. Grid Search找n=5(平滑天數), n1(已經fix=30), n2=5(預測幾天後)：在grid search後，我們決定採取smoothing與區段標準化，讓資料不至於因過往與現今的價差過大而使標準差過於特異。並將我們愈猜測的目標定為smoothing完5天後的價位。(5,10,15,20天）

4. train test validation 切割：我們採用8:1:1的情況來進行切割

C. 模型設計與實驗流程：

1. 模型的選取：在思考股價主要資料型態是偏向時續性資料後，我們決定使用傳統的LSTM來進行模型優化，並在中間插入batch normalization(詳細模型請見colab file)
2. 以accuracy來判斷整體猜測的準度，但因為即使準度夠高，也可能因為現實中的小賺大賠，致使我們無法得到獲利，所以我們第一步：先以訓練好的模型，對為各支成分股進行validation資料其間的獲利績效評估，以「投資報酬率」挑選出其中表現較為良好的幾支股票，以其獲利作為我們實際測試test的資產權重。
3. 說明我們validation計算的細節：例如可以買零股，每支股票分配一百萬的資金。

P.S. 科技部計畫：比較CNN，LSTM，RNN各模型的表現（預測的準確度及投資報酬率），並說明為何最後只選LSTM。

D. 模型結果：

1. 說明Training可達到0.88的Accuracy。訓練圖表及準確度
2. validation的投資報酬率
3. testing的投資報酬率

E. 結論：

1. 我們的模型讚，可以躺著賺。