資科FinalProject 統計112林家同

(3)Finmind是一個由台灣創作者開發的金融大數據資源包，這個資源包提供以台股為主的金融數據，使用者只要先在終端機安裝finmind，並申辦帳號，就可以取得台股技術面、基本面、籌碼面的data，像是個股的每日成交價、融資融卷數、本益比、股價淨值比等資訊，另外他們也有自己的Github，裡面有原始爬蟲的py檔，大家如果有興趣也可以去查看。

(4)(5)這個投影片將解釋如何取得資料，在jupyter輸入帳密後，會獲得一組token，我們可以把它想像成一個金鑰，拿著他便可連上finmind。接著設定完想要取得的個股和時間區段，就可以送出請求了，我們拿到的第一手資料都是一個完整的dataframe。分次取得技術面、基本面、籌碼面以及大盤的資料我們再將得到的多個dataframe做整合。整合後的dataframe包含了同一個股的多個金融數據，而這些特徵值就是column names。

(6)將資料整合後，下一步就可以開始處理多變量分析的問題，大家都知道交易市場裡充斥著影響股價的各種因子，所以我們的目標是將先前取得的特徵作為自變數，探討多項變數對收盤價y的影響。我們篩選特徵值的依據是用各個特徵的移動平均和收盤價算出相關係數的平均，利用python內建的函式rolling，再drop掉其中低度相關與負相關，經篩選後我們留下來13個特徵值，並建立新的dataframe。

(7)獲取一個要真正拿來訓練模型的dataframe後，我們要先對資料進行預處理，這邊運用的是將所有的資料進行(0,1)的標準化，消除特徵間單位和尺度，可以有助於演算法的優化。

(8)(9)因為我們拿到的資料是一個連續性的數量變化、隨時間先後排列的時間序列資料，所以在機器學習方法可以被使用之前，我們必須將時間序列重新建構成監督式學習問題，讓一個特徵值的序列變為一對序列的輸入和輸出。我們運用pandas中的shift()函示便可將序列進行一步兩步甚至是多步的時間平移，我們考慮市場十天前到當天的交易資訊會對當天的交易價做影響，所以透過數據平移的方式將我們的目標便量往前平移十個時間單位，建立13\*10+13的的新dataframe，有了標籤後它就變成了監督式學習的數據，這樣我們就可以訓練lstm模型了。

(15)對於一個連續性的數值，評估其最佳化模型有一個準則，利用回歸問題的損失函數，以下的兩張圖，x軸代表訓練的時間，y軸代表計算的損失。這裡將提供兩個常用的方法，MSE和MAE。相信大家對於MSE並不陌生，從上方的數學式可以看出，MSE就是計算實際值與預測值的平方差，再平均，而因為有平方項，確保Loss Function≥0，因此模型一定會向最小誤差的方向去收斂，然而outlier在MSE會被指數性的放大。而相對於MSE，MAE對outlier較為寬容，但它有一個令人詬病的缺點就是在接近最佳解時，MAE的部分梯度仍保持一個定值時。從右邊那張圖可以看出MAE無法在低損失時收斂，這樣的表現不利於模型的學習。另外，我們也可以適度引用判定係數，不管在任何模型，R-square都會被計算成0~1之間的分數，越接近1代表表現越好。Skleran還有另一個函示explained variance score,其計算方式與判定係數稍有不同，如果error的平均值為0，結果才會一樣。

(16) 前面有介紹過LSTM的每個神經元都有自己的權重，而在上一張投影片提及的損失函數，其實是利用權重w和bias計算出的預測值和真實值的函數，而優化器的任務，我們可以先粗解地理解為對權重和bias的調節。以下三個優化器由Tensorflow官方提供的，其中以AdamOptimizer最常被使用，因為它對學習率有個約束，使得每一次學習率都有一個確定範圍，這個應算法能快速找到參數更新並找到梯度正確下降的方向，而另外兩個優化器在收斂表現就不如adam.

(17) 在這個投影片中，我們提供LSTM模型參數輸出讓大家參考，左圖的輸出是建立在只有一層LSTM下，因此最後的output shape是一個二維數據；右邊的數據是建立在多層的LSTM，我們會在第一層LSTM加上returnsequence=True讓輸出為三維，以便進行LSTM的堆疊，這邊要注意的是最後的dense輸出一樣是二維的數據。因為有深層且分層的神經網路模型，所以不管在loss值及判定係數上表現都比單層還要好，但是在收斂速度上好像顯得過快，我們懷疑隱藏層的增加是否還要搭配更多的參數調整，不然會過度擬和的問題而導致梯度下降過快。但基本上,堆疊lstm在深度學習上能使模型更深入，幫我們預測更有挑戰性複雜的問題。

(18)以下我們將提供這次實作的結論，在這個投影片之前的測試，我們先是選定長榮公司的資料進行模型訓練，而得到相關的實驗數據。因此我們好奇，同樣lstm的baseline是否適用於其他公司。我們利用四家上市公司的資料做測試，並在畫面下方分別繪出過去五年的歷史股價。上方是損失函數收斂的情形，我們認為，台積電和聯發科的損失函數之所以能平穩的收斂，是因為他們的公司有較穩定的獲利能力，而使股價平穩的成長。因此在每輪lstm抽樣進行後，繪製出的擬和曲線之趨勢也較為一致，在loss function中較少出現收斂反升的情形。