資料分析 HW3-1 Report

F74056085 林哲寬

Q: How do you preprocess the dataset?

A:

1. 由於各種特徵的尺度不一，在使用之前進行了正規化，即減去平均後再除以標準差，以避免模型產生太大的偏差。事實上不進行正規化的話，我所使用的類神經模型是完全不堪使用的。

2. 增加了和上一筆記錄相比是增加還是減少的屬性。除了作為輸出的 label 之外，在處理時序性的資料時，這樣一個二元的特徵的確是有幫助的。

3. 定義數值的區間：因為連續性的數值無法直接餵給 Classifier 作運算，必須定義數個區間並用 LabelEncoder 再次轉換成數值。這邊使用 qcut 來實作，一共切成五等份。

Q: Which classifier reaches the highest classification accuracy in this dataset ?Why?

A:

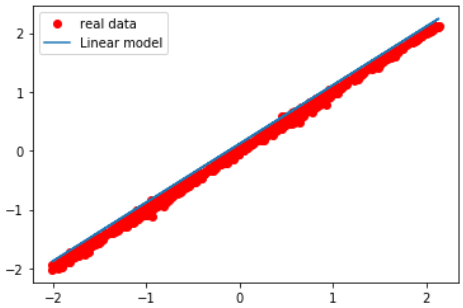
這次作業中我實作了三個分類器，分別是：

**1.Logistic Regression**

單獨使用了開盤價格這個屬性對股市走向作邏輯性的回歸，最終在這個 dataset 獲得的準確度是 **0.548**，只比隨便猜好一點而已，或許是特徵不夠強力或數目過少。

**2.Neural Network**

僅使用一層簡單的線性模型，一樣是利用正規化後的開盤價格迴歸出正規化後的收盤價格，與前一次的迴歸結果作比較藉此判斷出股市走向是上升或是下降。如下面一張圖即為迴歸出來的結果，看起來很準但實際上獲得的準確度只有 **0.52** 而已，或許是因為前一次和後一次的收盤價格差距通常都不大，就算只預測錯一點點，也有可能歸納如完全相反的結果。




**3.Random Forest Model**

這邊用上了全部的屬性來建立決策樹，最後得到的準確度是 **0.58**，是這三者之中最好的。這證明了股市的走向並不單純取決於單一屬性，即便他們的相關性看起來很高。

Q: Can this result remain if the dataset is different ?

A:

Logistic Regression 的 test accuracy 是 **0.528**，另一個資料集的 test accuracy 是 **0.595**

Neural Network 的 test accuracy 是 **0.627**，另一個資料集的 test accuracy 是 **0.647**

Logistic Regression 的 test accuracy 是 **0.631**，另一個資料集的 test accuracy 是 **0.595**

三者的表現變成是 Neural Network 最好，應該是該模型所擷取的特徵在這兩個資料集更為明顯的緣故。

Q: How did you improve your classifiers ?

A:

關於這個部分我修改了兩個模型，分別是：

1. Neural Network

改進方法：除了把 epoch 從 1000 加到 2000，讓 loss 可以降到更低之外，通過實驗我使用 low price 這個屬性取代原本的 open price 屬性，test accuracy 即提升到 0.806，另一個資料集則提升到 0.802。是迴歸的狀況變得更好，進而帶動準確率的提升。

2. Random Forest Classifier

改進方法：這個模型進了大幅度的改造，區間替換了成和上一個值相比，是上升還是下降的二元值，用趨勢來預測趨勢。這樣做之後模型的準確率要很大的提升，來到了 0.817，另一個資料集則提升到 0.849。說明這種改法十分有效。