資料科學導論
Competition 2

組員：蕭捷晨、黃書璿、施其均

壹、資料分析

(一)讀取資料與時間格式轉換

for 迴圈遍歷 L1_Train.csv 至 L17_Train.csv，接著利用 pandas 讀取檔案內容到 DataFrame，同時時間列 DateTime 被轉換為 datetime64 格式。

(二)篩選符合條件的資料

all_dates 是提取的所有日期、valid_hours 篩選出 9:00 至 17:00 時間段內的資料、valid_dates 取得有 9:00 至 17:00 資料的日期集合、no_data_date 找到沒有該時間段資料的日期集合、has_data_dates 找到有該時間段資料的日期集合。

(三)分類

no_data_df 是沒有 9:00 至 17:00 資料的資料，而 has_data_df 篩選出有 9:00 至 17:00 資料的資料。

(四)平均

將風速、氣壓、溫度、濕度、陽光和功率按每 10 分鐘計算平均值，並存於 numpy 陣列。

(五)預處理

依照 DataTime 將資料分組，接著篩選數據 before_9am 是當日早於 9 點的數據、after_9am: 則是當日 9:00 至 17:00 的數據。爾後，從 9 點前的數據中，向前批次取樣，最多計算 12 個批次的平均值，加上 0 的標籤，存儲於 before_9am_avg，再用 after_9am_avg 表示每 10 筆數據一組計算平均值，加上 1 的標籤，最多保留 48 個批次。

貳、方法嘗試。

(一)讀取 upload

擷取出 upload.csv 裡代表日期跟機器的編號，以便後續進行填答和輸入模型。

(二)過濾指定日期

讀取由 upload.csv 生成的 output.csv，由 Tag_Part 分組，對應日期存入列表。提取 DateTime，並轉為日期字串，然後比對是否屬於目標日期，接著將將符合條件的數據存入 temp_data，按日期分類。最後跟前處理一樣，提取 09:00:00 之前的數據，按時間降序排列，每 10 分鐘為一組，計算各指標的平均值，並存入 interval_data，再把將每組平均值計算結果存入 averaged_intervals，且標註來源標籤。

(三)資料集分割

將剛剛前處裡過的資料進行標準化，使用 MinMaxScaler 將輸入與輸出特徵縮放至 [0, 1]。input_data 是選擇 Label == 0 的資料作為輸入特徵，output_data 選擇 Label == 1 的資料作為目標特徵。接著把輸入數據按 12 天切片，然後對應的輸出數據按 prediction_horizon 生成。爾後，把數據的 80%用於訓練，20%用於測試。

(四)模型建構

建立 LSTM 模型，第一層 LSTM 有 128 個 Unit，第二層有 64 個，添加 Output 層防止過擬合，Dense 層 48 個時間的預測值。接著使 epoch=100 表示訓練 100 個迭代，每次更新模型都使用 320 個樣本，最後用測試集檢驗模型的表現。將訓練好的 17 個模型存為 h5 檔。

參、結果觀察

(一)加載模型

使用 load_model 函數讀取已訓練好的 LSTM 模型，增加 batch 維度，讓數據形狀符合批次大小、時間步、特徵數)，接著對每個批次進行預測，輸出標準化的目標值，最後將預測結果轉回原始單位

(二)展平結果並保存

使用 np.vstack 將所有批次的預測結果合併為一個完整的數據集，將預測結果保存為 no_data_L1_predictions.csv。

肆、心得

藉由此次 Competition 將綠能與模型訓練結合，我們體會到原來模型預測對生活的幫助大有裨益，透過早上九時前的資料即可以推斷當天九時後所缺失的資料。在實作中，我們發現資料的前處理對於結果的影響是舉足輕重的，經過合理前處理的資料，不僅使模型訓練時間大幅縮短，也使得輸出的誤差更小。同時，我們也學到新的模型的應用，不論是之前的 knn、sklearn 亦或是此作品使用的 LSTM 都是瑕瑜互見的模型訓練方式，尤其是 LSTM 這個長短期訓練，讓我們見識到模型訓練對長期短期時序的預測可以有多擬真。透過這次實作，我們領略到選取對的訓練方式是至關重要的，在長短期資料預測中，選擇用 LSTM 輔以線性回歸及標準化，相較於其他的訓練方式，其所展現的結果更具優勢，也更加貼合實際情形。