Machine Learning Week 4 Programming Assignment Report

111652042 顏友君

問題回顧：

(1) 將原始資料轉換為兩個監督式學習資料集：

(a) 分類 (Classification) 資料集

將溫度觀測值分類成有效(label = 1)及無效(label = 0)。

(b) 回歸 (Regression) 資料集

保留有效的溫度觀測值(剔除所有無效值)。

(2) 模型訓練

使用 (1) 中整理出的兩個資料集，分別訓練一個簡單的機器學習模型：

分類模型 (classification model)：

以 (經度, 緯度) 預測格點資料是否為有效值 (0 或 1)。

回歸模型 (regression model)：

以 (經度, 緯度) 預測對應的溫度觀測值。

模型說明：

我使用的模型是**隨機森林(random forest)**，剛開始有嘗試過使用較簡單的邏輯迴歸(logistic regression)來做分類，但可能是因為氣溫的分布本身就是非線性的，所以用這種線性回歸的模型來做效果就會比較差，準確率大概都落在 0.57 左右，而且氣溫的分布是被限制在一個範圍內，這種情況就比較適合利用決策樹來做分類，而隨機森林就是利用多個決策樹來做預測，於是就嘗試用隨機森林來做，而結果也表現得不錯，於是就將隨機森林保留下來當作本次作業的模型。

訓練過程：

分類模型及回歸模型皆使用 scikit learn 套件中的 random forest 類別，將 80%的資料當作訓練集，剩下 20%當作測試集，設定決策樹數量為 100 棵，最大深度為 10 層，再利用.fit 函式做訓練。

結果分析：

Grid shape: (120, 67)

總格點數：8040

分類結果

有效值數量：3495

無效值數量：4545

準確率：0.9857

- 視覺化分析一：

  將有效值與無效值的分佈畫出來，可以發現有效值都集中在陸地。

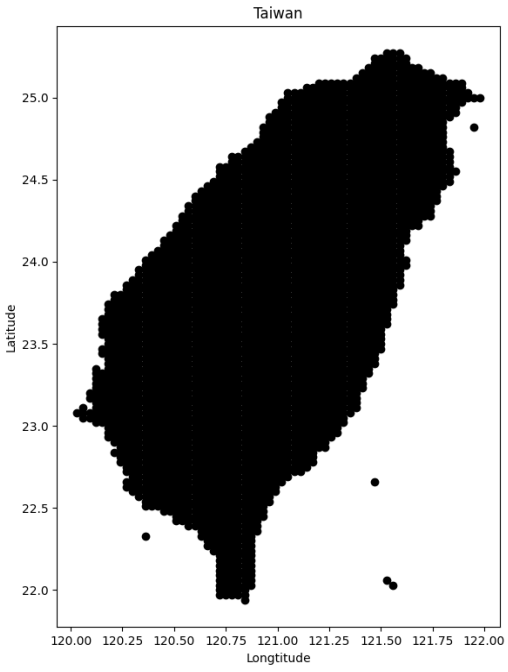

Fig.1 溫度觀測值分類後分布圖

- 視覺化分析二：

混淆矩陣(Confusion matrix)，將我們的預測結果分為以下四種類型：

		預測值(Pred)	
		0 (無效值)	1 (有效值)
真實值(True)	0 (無效值)	TP 預測正確	FP 預測錯誤
	1 (有效值)	TN 預測錯誤	FN 預測正確

實際預測結果(藍色越多，紅色越少)：

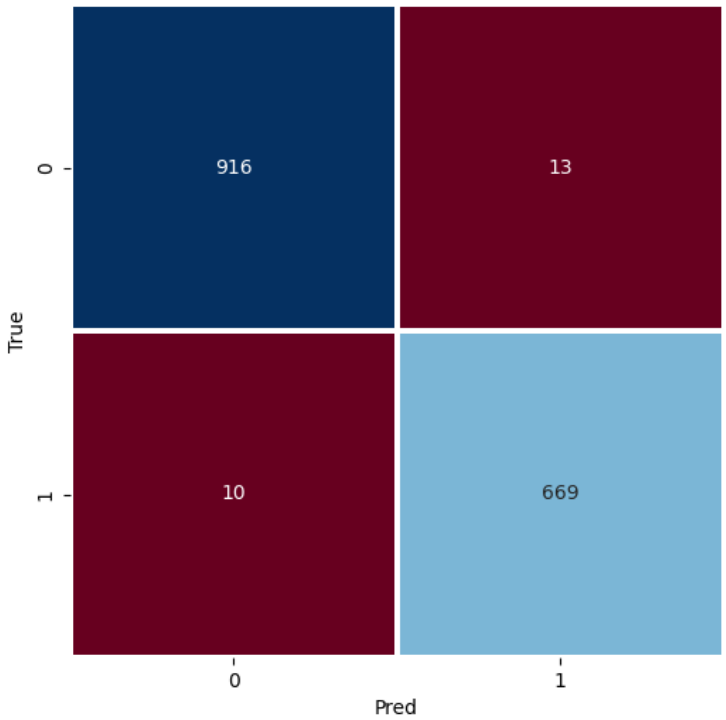

Fig.2 分類結果混淆矩陣

透過混淆矩陣可以看出，若將有效值及無效值分開來看，也是大部分都有判斷正確。

回歸結果

RMSE: 2.6509 °C

MAE: 1.7351 °C

R^2: 0.7928

透過以上三個數據可以看出，這個模型雖然能夠判斷出接近八成的溫度變化，但平均誤差(MSE)還是高達 1.74°C，且均方根誤差更高達 2.65°C，表示可能在某些溫度變化較劇烈的地方，像是山上，就會出現較大的離群錯誤。

結論：

在做本次作業的過程中額外收穫了很多新知識，像是面對不同種類的資料集，當原先的結果不如預期時，不妨換另一種模型試試看，每一種模型都有較適合訓練與較不適合訓練的資料型態，像這次換了一個模型，準確率就大大提升，另外，這次的預測都是從單一點資料來做判斷，但在實際氣象學的領域中，鄰近資料點的那些資料也是很重要的參考依據，或許下次有機會再做更進階的訓練再將那些也考量進去。