Programming Assignment 4

1. 三個 cell 分別的主要內容

Cell 1 — 解析 XML → 產生兩個 CSV

Cell 2 — 分類模型

- 前處理：data:(lon, lat)標準化
- 切分：60%/20%/20%
- 模型：MLP **2→16→32→1**，ReLU 激活，輸出 logits
- 訓練：`BCEWithLogitsLoss` + Adam(lr=1e-3)，batch=256，epoch=25
- 輸出：`sigmoid(logits)` → 機率 (0~1)。
- 評估：AUC-ROC、classification report、confusion matrix；繪製 ROC、機率熱度圖、機率分佈直方圖。
- 對應圖：【圖1】～【圖4】

Cell 3 — 回歸模型

- 前處理：僅有效值；`X=(lon,lat)` 與 `y=value` 各自標準化
- 切分：70%/15%/15%。
- 模型：MLP **2→16→16→1**，ReLU 激活，輸出為實數
- 訓練：`MSELoss` + Adam(lr=1e-3)，batch=256，epoch=30。
- 評估：MAE、RMSE；視覺化預測曲面與誤差直方圖。
- 對應圖：【圖5】～【圖7】

---

2. Classification Model

Data

- domain : $\mathcal{X} = \mathbb{R}^2$, $x = (\text{lon}, \text{lat})$
- codomain : $\mathcal{Y} = \{0, 1\}$
- data set : 8040 筆，60/20/20 分割（特徵已標準化）

Hypothesis（MLP）

$$h_\theta(x) = W_3\,\sigma\big(W_2\,\sigma(W_1 x + b_1) + b_2\big) + b_3, \qquad \hat{p}_\theta(y = 1 \mid x) = \frac{1}{1 + e^{-h_\theta(x)}}$$

其中 $\sigma = \mathrm{ReLU}$，層寬 $2 \to 16 \to 32 \to 1$。

Loss function（BCE with logits）

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log \hat{p}_\theta(x_i) + (1-y_i)\log\big(1-\hat{p}_\theta(x_i)\big)\Big].$$

結果分析（對應圖：1～4）

- **整體表現**：AUC=0.9952，Accuracy=0.9664。【圖1】【圖2】
- **分類指標**（test）：
  - class 0：P=0.9672、R=0.9736、F1=0.9704（n=909）
  - class 1：P=0.9654、R=0.9571、F1=0.9612（n=699）
- **混淆矩陣**（rows=true, cols=pred）：$\begin{bmatrix}885 & 24\\ 30 & 669\end{bmatrix}$ → FP=24、FN=30。【圖1】
- **可視化**：ROC 線貼左上（排序性極佳）【圖2】；機率熱度圖呈中央高、邊緣低的空間結構【圖3】；兩類機率分佈幾乎分離，重疊很少【圖4】。
- **訓練穩定性**：train/val loss 最終約 0.14/0.136，貼近、無明顯 overfitting。【圖1】

---

3. Regression Model

Data

- domain：$\mathcal{X} = \mathbb{R}^2$，$x = (\mathrm{lon}, \mathrm{lat})$
- codomain：$y \in \mathbb{R}$（溫度 °C），移除 −999 後保留有效樣本
- 標準化：$x' = \mathrm{Std}(x)$、$y' = \mathrm{Std}(y)$；標準化成 °C

Hypothesis（MLP）

$$\hat{y}_\theta(x) = W_3\,\sigma\big(W_2\,\sigma(W_1 x + b_1) + b_2\big) + b_3, \qquad 2 \to 16 \to 16 \to 1,\ \sigma = \mathrm{ReLU}.$$

Loss function（MSE）

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\big(\hat{y}_\theta(x_i') - y_i'\big)^2.$$

結果分析（對應圖：5～7）

- **訓練/驗證曲線**：MSE 由 $\sim 1.03 \to 0.45$（train）與 $\sim 0.99 \to 0.47$（val）穩定下降，無嚴重過擬合。【圖5】
- **測試表現**：MAE=$3.09^\circ\mathrm{C}$、RMSE=$4.12^\circ\mathrm{C}$（RMSE>MAE 代表少量大誤差）。【圖5】

- **預測曲面**：捕捉中央偏低、四周偏高的大尺度趨勢；邊界多邊形為凸包插值效果。【圖6】
- **誤差分佈**：多數殘差在 $[-5, 5]^\circ\mathrm{C}$，右尾較厚，少量 $> 10^\circ\mathrm{C}$ 的離群誤差拉高 RMSE。【圖7】
- **改進方向**：加入海拔/距海/鄰域統計等domain裡的變數；改用 Huber/MAE 損失、學習率衰減與早停；比較 GBDT/GPR/RBF 或在 MLP 中加 BatchNorm/Dropout。

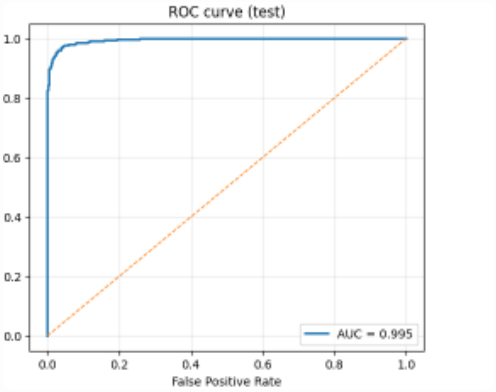
---

4. 參考圖

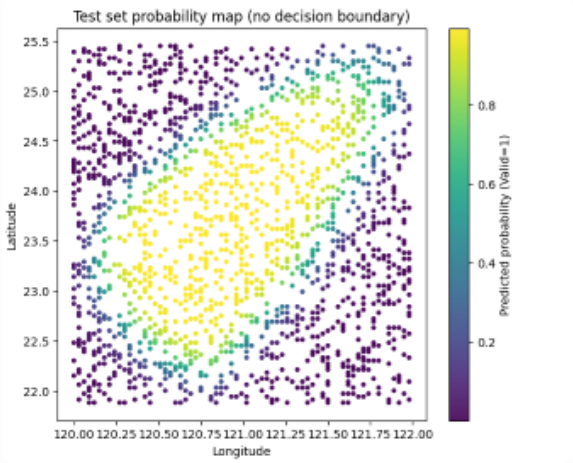
- 【圖1】Classification train

```
[01] train loss=0.6735 | val loss=0.6611
[05] train loss=0.5572 | val loss=0.5364
[10] train loss=0.3471 | val loss=0.3321
[15] train loss=0.2250 | val loss=0.2182
[20] train loss=0.1693 | val loss=0.1636
[25] train loss=0.1412 | val loss=0.1357

Test AUC = 0.9952

Classification report (test):
              precision    recall  f1-score   support

         0.0     0.9672    0.9736    0.9704       909
         1.0     0.9654    0.9571    0.9612       699

    accuracy                         0.9664      1608
   macro avg     0.9663    0.9653    0.9658      1608
weighted avg     0.9664    0.9664    0.9664      1608

Confusion matrix (test) [rows=true, cols=pred]:
[[885  24]
 [ 30 669]]
```

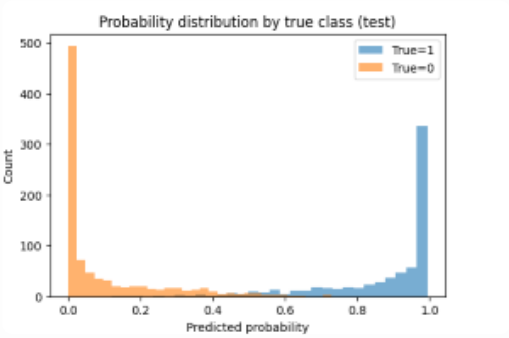
- 【圖2】ROC curve (test)



- 【圖3】Test set probability map



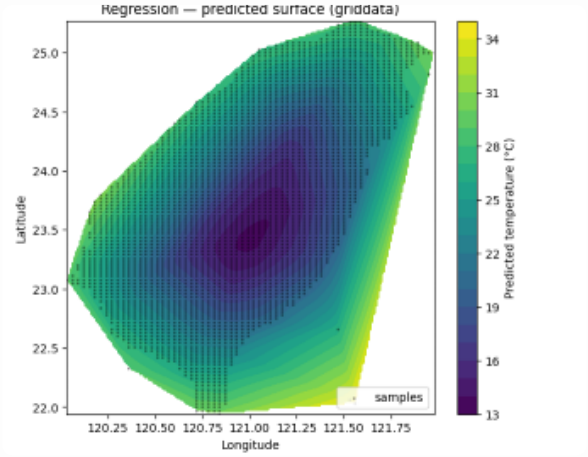
- 【圖4】Probability distribution by true class



- 【圖5】Regression test and train

```
[01] train MSE=1.0345 | val MSE=0.9930
[05] train MSE=0.9285 | val MSE=0.8926
[10] train MSE=0.7661 | val MSE=0.7433
[15] train MSE=0.6323 | val MSE=0.6348
[20] train MSE=0.5436 | val MSE=0.5612
[25] train MSE=0.4881 | val MSE=0.5117
[30] train MSE=0.4458 | val MSE=0.4726
Test MAE = 3.091 °C
Test RMSE = 4.121 °C
```

- 【圖6】Regression — predicted surface (griddata)



- 【圖7】Test error histogram (pred−true, °C)