ProgrammingAssignmentReport_111652026劉馥瑞

摘要

本報告旨在透過結合分類模型 C(x) 與迴歸模型 R(x)，建構一個分段平滑函數 h(x)，用於處理氣溫網格資料中的無效值。我們使用二次判別分析 (GDA/QDA) 作為 C(x) 來預測資料的有效性，並使用線性迴歸作為 R(x) 來預測有效區域的氣溫。結果顯示，GDA 在分類有效數據區域方面準確度為 83.02%，且最終建構的 h(x) 函數成功地根據 C(x) 的預測，將輸出值平滑地分段為預測溫度或無效值。

一、資料處理與模型架構

1.1 資料處理與標籤化

原始氣溫網格資料（XML 檔案）包含 8040 個格點。我們將每個格點的經緯度 x=(Lon,Lat) 作為特徵，氣溫值作為目標變數 y。根據作業要求，數據被分為兩類：

- 分類標籤 Ycls：
  - Ycls=1（有效數據）：氣溫 =−999.0。數量：3495。
  - Ycls=0（無效數據）：氣溫 =−999.0。數量：4545。
- 迴歸數據集：僅包含 Ycls=1 的有效經緯度特徵 Xreg 及氣溫值 Yreg。

1.2 分段平滑函數 **h(x)** 定義

最終目標是定義以下分段函數 h(x)：

$$h(x)=\{R(x)-999.0 \text{if } C(x)=1 \text{if } C(x)=0$$

其中：

- C(x): 採用自定義的 GDA 分類器（QDA 形式），用於判斷格點 x 是否位於有效數據區域（即 C(x)=1）或無效區域（即 C(x)=0）。
- R(x): 採用在有效數據集 Xreg,Yreg 上訓練的線性迴歸模型，用於在有效區域內進行溫度預測。

1.3 組合函數 **h(x)** 建構方法

h(x) 模型的建構採用了「先分類，後迴歸」的串聯邏輯，旨在解決氣象網格數據中有效數據與無效數據的混合問題。

建構流程概述：

1. 資料分流與訓練集劃分：數據集首先被劃分為兩組：
   - 分類訓練集：包含所有格點 (8040 筆)，用於訓練 GDA 模型 C(x)。
   - 迴歸訓練集：僅包含有效溫度觀測值 (3495 筆)，用於訓練線性迴歸模型 R(x)。
2. 模型訓練：分別獨立訓練 GDA C(x) 和線性迴歸 R(x)。

3. 條件式組合：最終的 h(x) 函數在程式碼中實現為一個條件判斷邏輯：對於任何新的經緯度輸入 x，程式會先呼叫 C(x) 進行判斷。
    - 若 C(x) 輸出為 1 (有效)，則計算 R(x) 的預測值作為 h(x) 的輸出。
    - 若 C(x) 輸出為 0 (無效)，則直接輸出恆定值 −999.0。

這種方法確保了迴歸模型 R(x) 僅在 GDA 確信為有效數據的地理區域內運行，從而在 C(x)=1 區域內提供了平滑的溫度場。

二、模型訓練與性能分析

2.1 GDA 模型工作原理與適用性

高斯判別分析（GDA）是一種生成式的分類演算法，其核心假設是：每個類別的數據點都是由一個多元高斯（常態）分佈產生的。

工作原理：GDA 的訓練過程是估計每個類別 k 的高斯分佈參數：先驗機率 ϕ_k、平均向量 μ_k（分佈中心）和協方差矩陣 Σ_k（分佈形狀）。分類時，GDA 利用貝氏定理來計算新數據點 x 屬於類別 k 的後驗機率 $P(Y=k|x)$：

$$P(Y=k|x) \propto P(x|Y=k)P(Y=k)$$

GDA 將 x 分類到使這個乘積最大的類別。

- $P(x|Y=k)$ 是似然機率，由類別 k 的高斯分佈密度函數計算得到。
- $P(Y=k)$ 是先驗機率 ϕ_k。

GDA 根據對協方差矩陣的處理方式分為：

- LDA (線性判別分析)：假設所有類別共享一個協方差矩陣 Σ，產生線性決策邊界。
- QDA (二次判別分析)：允許每個類別有獨立的協方差矩陣 Σ_k（本任務採用此形式），產生非線性決策邊界。

適用於此數據集的原因：此氣象分類任務旨在區分台灣陸地的有效數據 (C=1) 與海域/山區的無效數據 (C=0)。

1. 需要非線性邊界：台灣島的地理輪廓複雜且彎曲，分類邊界不是簡單的直線。如 Section 2.3 所示，QDA 形式下非線性的決策邊界，提供了顯著優於線性模型（Logistic Regression）的準確度，能夠更好地擬合實際的地理邊界。
2. 分佈形狀差異大：陸地（有效數據）的分佈相對緊湊，而海域（無效數據）的分佈更為分散。QDA 允許兩類數據擁有獨立的協方差矩陣 Σ_k，能夠精確地建模這些形狀和方向差異巨大的空間分佈。

2.2 GDA 分類模型 **C(x)** 性能報告

GDA 模型旨在模擬兩類數據（有效 C=1 與無效 C=0）的多元高斯分佈，並使用獨立的協方差矩陣 Σ_k（即 QDA 形式）。

- 訓練/測試集劃分：使用 80% 訓練，20% 測試。
- GDA 測試集準確度：83.02%。

- 基線模型（Logistic Regression）測試集準確度：56.53%。

分析：GDA 分類器以顯著的優勢超越了簡單的 Logistic Regression 模型（83.02% vs 56.53%），這表明有效數據與無效數據的分佈在經緯度特徵空間中呈現出明顯的非線性（或非共線）結構，非常適合使用 QDA 這類更複雜的二次決策邊界來劃分。圖表 1（左圖）中的橙色虛線決策邊界即是 GDA 成功的證明。

2.3 迴歸模型 **R(x)**

R(x) 訓練於所有 3495 個有效溫度觀測值上。其結果如下：

- R(x) 迴歸模型係數：w=[−4.58023674,2.53346728]（分別對應經度和緯度）。
- 截距：b=515.51。

這表明：

1. 在控制緯度不變的情況下，經度每增加一度，氣溫約下降 4.58◦C（由於經度係數為負）。
2. 在控制經度不變的情況下，緯度每增加一度，氣溫約上升 2.53◦C（由於緯度係數為正）。注意：這些係數的解釋性較弱，因為經緯度的絕對數值較大，截距 515.51 很高。該線性模型僅用於在 GDA 確定的有效區域內進行「平滑」的溫度估計。

三、分段函數 **h(x)** 驗證與結果

3.1 函數行為驗證

通過對測試集中的六個點進行 h(x) 函數計算，驗證了其分段邏輯的正確性：

特徵 **(Lon, Lat)**	**GDA C(x)**	**R(x)** 預測值	**h(x)** 最終輸出	分段驗證
(121.95, 24.16)	0	18.1553	-999.0000	−999 成功應用
(121.50, 22.39)	0	15.7321	-999.0000	−999 成功應用
(121.59, 23.02)	0	16.9160	-999.0000	−999 成功應用
(121.11, 23.86)	1	21.2426	21.2426	R(x) 成功應用
(121.17, 24.49)	1	22.5639	22.5639	R(x) 成功應用

(120.78, 23.08)	1	20.7780	20.7780	R(x) 成功應用

驗證結果完全符合 h(x) 的分段定義：當 GDA 預測為無效區域 (C(x)=0) 時，輸出值為 −999.0；當 GDA 預測為有效區域 (C(x)=1) 時，輸出值為 R(x) 的迴歸預測值。

3.2 繪圖結果分析

Graph1(在Week_6資料夾中)（右圖）展示了 h(x) 函數在整個網格空間上的行為，明確達成了分段平滑的設計目標。

1. GDA 決策邊界可視化：Graph1（左圖）展示了訓練資料點的分佈，以及 GDA 模型通過等高線 (Level=0) 劃分的非線性決策邊界（橙色虛線）。這個邊界代表了後驗概率 $\log P(C=1|x) - \log P(C=0|x) = 0$ 的集合，成功地將主要集中在台灣陸地上的有效數據點 (C=1) 與集中在周邊海域和山區的無效數據點 (C=0) 分離。
2. h(x) 分段行為展示：Graph1（右圖）將 h(x) 的輸出映射到顏色上，其表現出明顯的兩段性：
    - 無效值區域（C(x)=0）：被 GDA 劃分為 −999.0 的格點，在圖上顯示為均勻的灰色區域。這有效地將氣象資料中的缺失值區域（如海面）與分析區域隔離開來。
    - 有效預測區域（C(x)=1）：在 GDA 邊界內的區域，顏色顯示為 R(x) 函數的輸出值。由於 R(x) 是一個線性迴歸模型，在該區域內，溫度等高線呈現出平滑且連續的變化，證明了 h(x) 在此區域內的平滑特性。
3. 邊界處的間斷：黑色實線代表 GDA 的決策邊界，它同時也是 h(x) 函數的間斷點。在該邊界上，函數值從 R(x) 的預測溫度值（例如 20◦C 左右）驟然跳變到 −999.0，清晰地體現了模型 h(x) 的分段定義。

四、結論

本專案成功地實現了一個強健的分段平滑函數 h(x)，它有效地解決了氣象網格資料中無效值 (T=−999.0) 的處理問題。透過將空間分類 (GDA) 與溫度預測迴歸 (Linear Regression) 結合，模型能夠：

1. 準確識別 氣溫觀測有效的地理區域 (C(x))。
2. 在這些有效區域內，提供平滑且可解釋的溫度估計 (R(x))。
3. 在無效區域，則明確輸出指定代碼 −999.0，避免了對缺失數據的無效內插。

最終的 h(x) 函數模型既能處理分類任務，又能執行迴歸預測，為後續的氣象資料分析和視覺化提供了一個乾淨且結構化的數據集。

註：本次報告及程式碼皆使用chatgpt作為輔助工具。