Machine Learning HW1

313513050 吳鈞皓

March 26, 2025

1 Introduction

本次作業使用 regression model 預測並重建南台灣地區的海拔高度資訊，用 MSE 作爲主要的評估方法，本次作業總共使用了三個 regression approaches ，分別是：

1. Maximum Likelihood (ML)
2. Maximum A Posteriori (MAP)
3. Bayesian Linear Regression

以下會對三者進行描述與分析。

2 Maximum Likelihood Approach

使用 Gaussian function 作爲 basis function 進行特徵轉換：

$$\phi_j(x_1, x_2) = \exp\left(-\frac{(x_1 - \mu_{j1})^2}{2\sigma_{j1}^2} - \frac{(x_2 - \mu_{j2})^2}{2\sigma_{j2}^2}\right)$$

2.1 程式邏輯：

(i) 將 $O_1 \times O_2$ 個 Gaussian basis functions 均勻分布在座標平面 (三維空間上的 xy-plane) 上。

(ii) 將 training data 的 xy 座標代入 basis function ，求得特徵矩陣 Φ。

(iii) 將 Φ 的 pseudo-inverse 乘上 training data 的 z 座標 (海拔) ，即可求得權重 w。

2.2 超參數設定：

(i) Basis 數量：$O_1 = O_2 = 25$ (共 625 個 basis)。

(ii) σ_{j1}, σ_{j2}：分別設爲 $1.5/O_1$, $1.5/O_2$ (可隨參數量動態調整)。

(iii) S-fold Cross Validation：4-fold 。

2.3 實驗結果：

從測試集得到的 MSE = 493.52 ，其中 MSE 的計算公式如下：

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(t_i - y(\mathbf{x}_i, \mathbf{w}))^2$$

以下是將測試集的座標資料拿來預測海拔高度所繪製而成的圖：

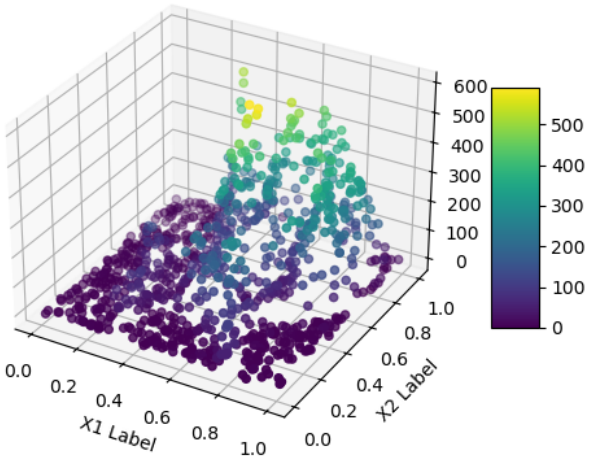

Figure 1: ML approach 預測的海拔資料。

3 MAP Approach

在 ML 方法的基礎上加入 regularization term 避免 over-fitting。
目標函數爲：

$$E_{MAP}(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{t} - \Phi\mathbf{w}\right\|^2 + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2$$

3.1 程式邏輯：

在 ML 方法求權重的公式中，加入 λI 作爲 regularization term：

$$\mathbf{w}_{MAP} = \left(\lambda\mathbf{I} + \Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t}$$

3.2 超參數設定：

(i) S-fold Cross Validation：5-fold (Basis, σ 同 ML)。

(ii) Regularization parameter λ：10^{-8}。

3.3 實驗結果：

從測試集得到的 MSE = 529.16 ，略微上升。
以下是將測試集的座標資料拿來預測海拔高度所繪製而成的圖：

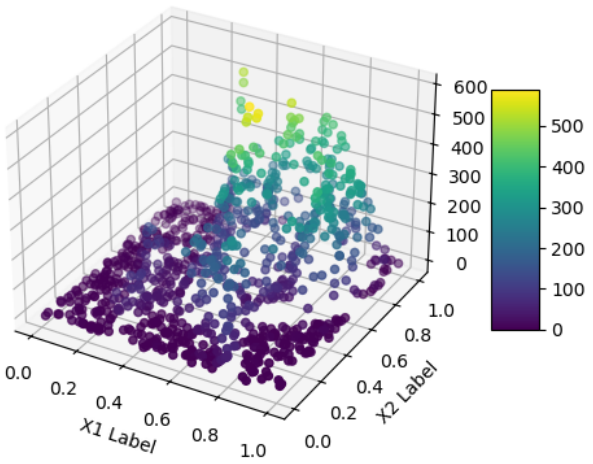

Figure 2: MAP approach 預測的海拔資料。

4 Bayesian Linear Regression Approach

本實驗使用了 Bayesian linear regression 方法來預測海拔高度，除了預測結果之外，本方法還可以描述預測結果的精確度，使我們對於預測結果有更多的了解。其中，prior 分佈的參數如下：

$$\mathbf{S}_0 = \alpha^{-1}\mathbf{I}, \quad \mathbf{m}_0 = \mathbf{0}$$

而以下則是 posterior 分佈的參數計算公式：

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^T\mathbf{t}\right)$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\Phi^T\Phi$$

4.1 程式邏輯：

(i) 簡化假設：$S_0^{-1} = \alpha I$, $m_0 = 0$。

(ii) 求解 posterior mean m_N 值作爲最終權重

4.2 超參數設定：

1. S-fold Cross Validation：5-fold (Basis, σ 同 ML)。
2. $\beta = 5$, $\alpha = 10^{-8} \times \beta$。

4.3 實驗結果：

從測試集得到的 MSE = 529.08 ，比起 MAP 方法略爲下降，但仍比 ML 方法高。

以下是將測試集的座標資料拿來預測海拔高度所繪製而成的圖：

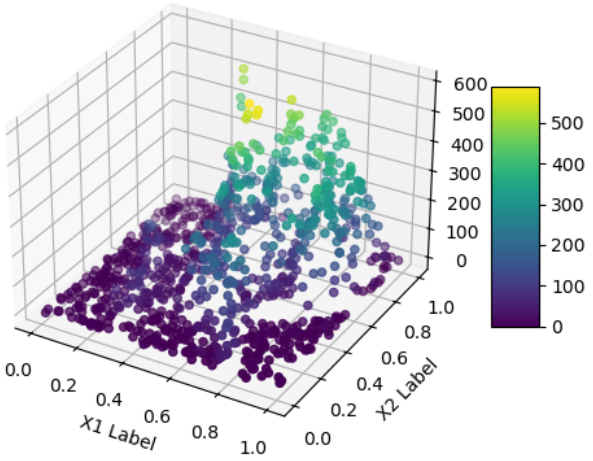

Figure 3: Bayesian linear regression approach 預測的海拔資料。

5 Disscussion

通常在較複雜的模型中，ML 方法因容易產生 over-fitting ，效果不如另外兩種方法，但由於 S-fold Cross Validation 的加入，over-fitting 的發生機會被有效的抑制，固後面兩種方法反而可能因爲複雜度較高，效果不如 ML 方法得出的結果。