Maximum Likelihood and Bayesian Linear Regression

從公式來看 Maximum Likelihood 在預測時他只會去找尋一個最優解的回歸式，但不會考慮到模型的整體複雜度，而 Bayesian Linear Regression 比 Maximum Likelihood 多考慮了 prior，在假設 prior 是高斯分布的狀況下可推得在樣本數增加後預測的變數減少，產生一個正則化的作用，而從程式的實驗解果來看，如果將正則化的 λ 設定為 1 並且 O1 和 O2 都設定為 2 去做最優化可發現 loss 值反而比 Maximum Likelihood 還低，但如果將 λ 設定為 1e-5 即可得到比 Maximum Likelihood 還優的解，而當 O1 和 O2 設定為 10 即可發現 Bayesian Linear Regression 的 loss 比 Maximum Likelihood 小超過一萬倍，因此推測在 O1 和 O2 較小的狀況整個模型並不複雜因此有無正則化的結果並不明顯，但假設 O1 和 O2 大的話，整體模型複雜的狀況有正則化的優點就十分明顯了。

解聯立方程式對比梯度下降

在我們列出公式後我們可以選擇使用解聯立方程式或是梯度下降法來求解最優的參數，梯度下降法可參考 hw2_SGD.py，解聯立方程式可參考 hw2.py，在 O1 和 O2 同為 2 的狀況下我們可以發現解聯立方程式可將 loss 降到 0.006 但梯度下降只能到 0.01，而梯度下降法中模型的參數以隨機均勻分布和 batch size 設定為 1 來講結果會最好，不過當 O1 和 O2 同為 5 的時候會發現解聯立方程式的 Maximum Likelihood 的 loss 到 0.4 而 Bayesian Linear Regression 為 0.007，但梯度下降法的結果均為 0.01，因此推測梯度下降法要到最優解的狀況比較難，但比較不容易陷入區域最佳解的狀況。

共同高斯分布

在一開始取 Feature Vector 時需要使用的 x 的 min 和 max，如果將測

試資料的 Feature Vector 裡面的 x 的 min 和 max 取的跟訓練時一樣可再將 loss 降低。如果取的是測試資料的 min 和 max 在 Maximum Likelihood 可以得到 0.008 而如果取一樣可以降到 0.006。

O1 和 O2 參數設置

我嘗試窮舉 O1 和 O2 在 2~15 所有組合，在解聯立方程式的方法下 Maximum Likelihood 的模型下可觀察到在某些特殊的位置有可能會導致 loss 暴增，而隨著 O1*O2 越大 Loss 值也越大，最小可以到達 0.006，最大可到達 7481，而在 Bayesian Linear Regression 也會有隨著 O1*O2 越大 loss 值也越大的趨勢，但不會發生突然的暴增，平均 loss 落在 0.007，可參考以下圖，xy 軸分別代表 O1 和 O2，y 代表 loss 值。因此選擇最優的 O1 和 O2 為 2。而使用梯度下降法也可觀測到類似遞增數據。