期中報告

R26094022 統計所 李權恩

1. Introduction

在本文當中，我們根據 Regularization Paths for Generalized Linear Models via Coordinate Descent 這篇文章中的兩個迴歸模型在我們的資料集上，兩個模型分別為 Lasso linear regression 以及 Lasso logistic regression。再利用此兩模型建立解釋變數與反應變數之間的關係後使用 AIC 及 BIC 兩種訊息準則來挑選適當的懲罰項係數λ。

在資料集上，主要分成 training 與 testing 兩個，資料筆數分別為 10000 筆與 23334 筆。變數上一個有 122 個，其中一個為反應變數有(A、B)兩類，實作上我們將此變數重新編碼為 1 與-1(A = 1, B = -1)或是 0 與 1(A = 1, B = 0)來使用，其他剩餘的 121 個變數作為解釋變數，在 121 個解釋變數當中有非常多的變數是隨機生成，與反應變數之間毫無關聯，而如果直接考慮使用所有的變數來進行模型建立，不但可能建立出錯誤的關聯性也容易有過度配適的問題發生。

而 Coordinate Descent 與傳統的 gradient decent 相比起來主要的特色在於 Coordinate Descent 在每次更新僅考慮搜尋單變量的最佳方向，也沒有考慮到學習率的問題，在迭代次數上較 gradient decent 來的少。

2. Methodology

下方分別描述了關於 lasso linear regression 及 lasso logistic regression 兩種方法的目標函數與參數迭代更新方式。而在使用梯度下降的相關方法中，我們可以考慮先統一所以解釋變數的單位，有了單位化的動作我們可以在額外獲得一些條件：

$$\sum_{i=1}^{N} x_{ik} = 0\,, \sum_{i=1}^{N} x_{ik}^2 = 1\,, \forall k = 1,2,\dots,p$$

有了上方的條件後更新算法可以做額外整理與簡化使迭代更加快速。

2.1. Lasso linear regression

在一般線性迴歸的預測任務中常使用均方誤差(MSE)來做為目標函數，希望能夠找到一組參數能使目標函數最小化，此外為了避免參數過多所以加入懲罰項$P_\alpha(\beta)$，所以目標如下：

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} R_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda P_\alpha(\beta)\right]$$

其中$P_\alpha(\beta)$是一項靠α來調整懲罰項，這裡寫的是一個通用的版本，而當α等於 1 時為 Lasso regression 反之當α等於 0 時為 Ridge regression，式子如下：

$$P_\alpha(\beta) = \frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|$$

在 Coordinate Descent 的方法上，每次迭代僅沿著其中一個維度的方向進行搜索來更新參數。考慮$\beta_j > 0$：

$$\left.\frac{\partial R}{\partial \beta_j}\right|_{\beta=\tilde{\beta}_{(-j)}} = -\frac{1}{N}\sum_{i=1}^{N} x_{ij}\left(y_i - \tilde{\beta}_0 - x_{i(-j)}^T\tilde{\beta}_{(-j)} - x_{ij}\beta_j\right) + \lambda(1-\alpha)\beta_j + \lambda\alpha$$

$$\tilde{\beta}_{(-j)} = \left(\beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_p\right)^{\mathrm{T}}$$

$$Set\ \left.\frac{\partial R}{\partial \beta_j}\right|_{\beta=\tilde{\beta}_{(-j)}} = 0$$

更新的方式為：

$$\tilde{\beta}_j \leftarrow \frac{S(\frac{1}{N}\sum_{i=1}^{N} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right), \lambda\alpha)}{1+\lambda(1-\alpha)}$$

其中S函數為：

$$S(z,\gamma) = sign(z)(|z|-\gamma)_+ = \begin{cases} z-\gamma \ \ if \ \ z>0 \ \ and \ \ \gamma<|z| \\ z+\gamma \ \ if \ \ z<0 \ \ and \ \ \gamma<|z| \\ 0 \ \ if \ \ \gamma>|z| \end{cases}$$

2.2. Lasso logistic regression

考慮反應變數為二元分類，例如$Y \in \{0,1\}$，在分類任務上羅吉斯迴歸是一種常見的方法，模型利用解釋變數來建立條件機率的關係：

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = \beta_0 + x^T\beta$$

或是也可以表示成：

$$\begin{aligned} P(Y=1|x) &= \frac{1}{1+\exp(-(\beta_0+x^T\beta))} \\ P(Y=0|x) &= 1-P(Y=1|x) \\ &= \frac{\exp(-(\beta_0+x^T\beta))}{1+\exp(-(\beta_0+x^T\beta))} \end{aligned}$$

而在這樣的分類任務上，我們的目標是設法讓 *log-likelihood* 函數值最大，但又要確保參數不會太多造成過度配適所以納入懲罰項，目標如下：

$$\max_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{N}\sum_{i=1}^{N}\left\{I(y_i=1)\log\big(p(x_i)\big) + I(y_i=0)\log(1-p(x_i))\right\} - \lambda P_\alpha(\beta)\right]$$

而上式在的 *log-likelihood* 的部分可以再整理為：

$$l(\beta_0,\beta) = \frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + x_i^T\beta) - \log(1+\exp(\beta_0+x_i^T\beta))$$

為了使參數的更新式更加容易呈現，所以將上方 *log-likelihood* 函數透過泰勒展開對$(\tilde{\beta}_0,\tilde{\beta})$做展開到二次式可得：

$$l_Q(\beta_0,\beta) = -\frac{1}{2N}\sum_{i=1}^{N} w_i(z_i-\beta_0-x_i^T\beta)^2 + C\big(\tilde{\beta}_0,\tilde{\beta}\big)^2$$

其中

$$\begin{aligned} z_i &= \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i-\tilde{p}(x_i)}{\tilde{p}(x_i)(1-\tilde{p}(x_i))} \\ w_i &= \tilde{p}(x_i)(1-\tilde{p}(x_i)) \\ & C\big(\tilde{\beta}_0,\tilde{\beta}\big)^2 \ \ \text{is constant} \end{aligned}$$

藉此我們的目標函數可以改寫為：

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} R_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[-l_Q(\beta_0,\beta) + \lambda P_\alpha(\beta)\right]$$

我們一樣對單變量做偏微分

$$\left.\frac{\partial R}{\partial \beta_j}\right|_{\beta=\tilde{\beta}_{(-j)}} = \frac{1}{N}\sum_{i=1}^{N} w_i\big(z_i - \tilde{\beta}_0 - x_{i(-j)}^T\tilde{\beta}_{(-j)} - x_{ij}\beta_j\big)\big(-x_{ij}\big) + \lambda(1-\alpha)\beta_j + \lambda\alpha$$

$$\tilde{\beta}_{(-j)} = \big(\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p\big)^{\mathrm{T}}$$

$$Set \ \ \left.\frac{\partial R}{\partial \beta_j}\right|_{\beta=\tilde{\beta}_{(-j)}} = 0$$

$$\left[\frac{1}{N}\sum_{i=1}^{N} w_i x_{ij}^2 + \lambda(1-\alpha)\right]\beta_j = \frac{1}{N}\sum_{i=1}^{N} w_i x_{ij}(z_i - \tilde{\beta}_0 - x_{i(-j)}^T \tilde{\beta}_{(-j)}) - \lambda\alpha$$

更新的方式為：

$$\tilde{\beta}_j \leftarrow \frac{S(\frac{1}{N}\sum_{i=1}^{N} w_i x_{ij}(z_i - \tilde{\beta}_0 - x_{i(-j)}^T \tilde{\beta}_{(-j)}), \lambda\alpha)}{\frac{1}{N}\sum_{i=1}^{N} w_i x_{ij}^2 + \lambda(1-\alpha)}$$

其中S函數為：

$$S(z,\gamma) = sign(z)(|z|-\gamma)_+ = \begin{cases} z-\gamma \ \ if \ \ z>0 \ \ and \ \ \gamma<|z| \\ z+\gamma \ \ if \ \ z<0 \ \ and \ \ \gamma<|z| \\ 0 \ \ if \ \ \gamma>|z| \end{cases}$$

3. Numerical result

上一節的參數更新裡都是考慮給定λ，而λ其實是一個需要做校調的參數，此節我們使用 BIC 訊息準則來選擇最合適的λ。下方為 BIC 訊息準則的公式：

$$\mathrm{BIC} = k\ln(n) - 2\ln(\hat{L})$$

其中 k 為估計的參數個數、n 為樣本數、$\hat{L}$為 likelihood 函數的最大值，藉由 BIC 準則可以同時衡量模型的解釋能力以及參數數量，BIC 代表的是一種相對關係，如果 BIC 越小代表模型相對精簡的同時也具備具備解釋能力。

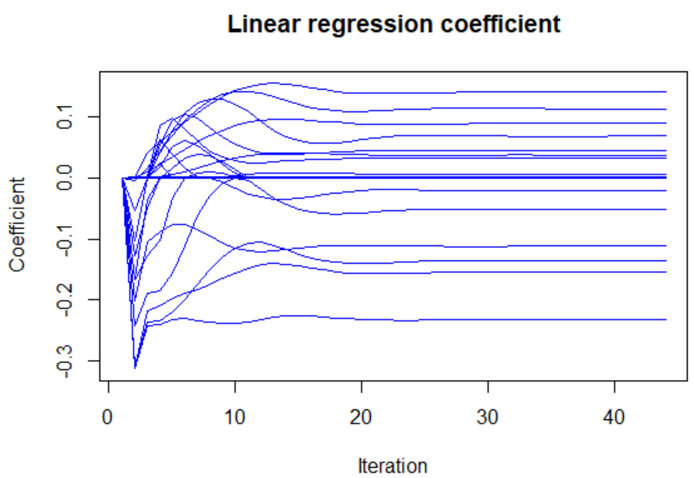
3.1. Lasso linear regression

透過 BIC 訊息準則我們選擇$\lambda = 0.02$，停止條件設定為$\left\|\tilde{\beta}_t - \tilde{\beta}_{t-1}\right\|^2 < 0.0001$，所得到的參數估計如下：

Active features	x_5	x_6	x_9	x_{10}	x_{11}	x_{12}	x_{13}
Coefficient	0.0324	0.0452	-0.0518	-0.1358	-0.1545	-0.2324	-0.1118

Active features	x_{14}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}
Coefficient	-0.0208	0.0690	0.1142	0.1412	0.0897	0.0381	0.0070

下圖為隨著迭代次數上升，各個參數的更新狀況，大約在迭代 20 至 30 次之間穩定。



而線性模型所預測的結果$\hat{y}$並非 1 或-1 的二元分類，所以我們設定$I(\hat{y}_i > 0)\ \forall i$為最後的預測分類。最後下表為 testing 資料的預測結果混淆矩陣，預測正確率為 92.14%。

混淆矩陣		True A	True B
Predict	A	10560	634
Predict	B	1200	10940

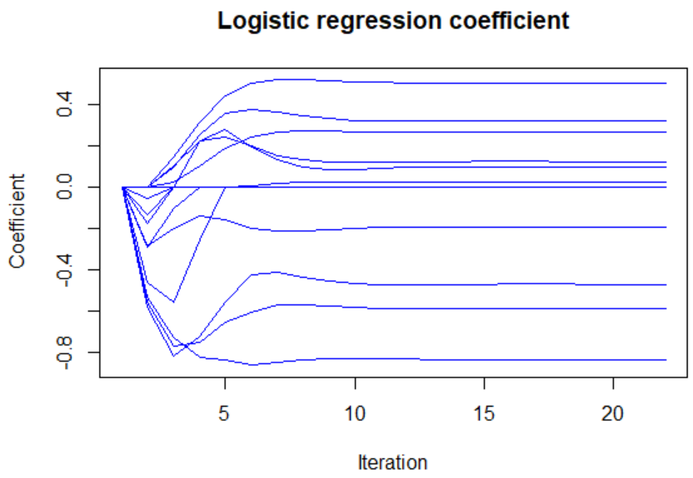
3.2. Lasso logistic regression

透過 BIC 訊息準則我們選擇$\lambda = 0.004$，停止條件設定為$\left\|\tilde{\beta}_t - \tilde{\beta}_{t-1}\right\|^2 < 0.0001$，所得到的參數估計如下：

Active features	x_{10}	x_{11}	x_{12}	x_{13}	x_{16}
Coefficient	-0.4691	-0.5862	-0.8325	-0.1952	0.0978

Active features	x_{17}	x_{18}	x_{19}	x_{20}
Coefficient	0.3244	0.5059	0.2659	0.0228

下圖為隨著迭代次數上升，各個參數的更新狀況，大約在迭代 5 至 10 次之間穩定。



考慮以 0.5 作為閥值(i.e.$I(\hat{y}_i > 0.5)\ \forall i$)，下表為在 test 資料上的分類結果混淆矩陣，預測正確率為 91.61%。

混淆矩陣		True A	True B
Predict	A	10510	707
Predict	B	1250	10867

4. Reference

1. Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software, January 2010, Volume 33, Issue 1.
2. Bayesian information criterion(BIC), wikipedia