最佳化方法 Optimization methods

期中報告 統計所 楊承翰 R26094064

1. 摘要

在一般的線性迴歸模型中，為了最佳化目標函數(最小化誤差平方和)，資料需符合許多假設，才能得到不偏迴歸係數，使模型變異量最低。對於實際資料而言可能有多個解釋變數，使得模型假設不成立與過度配適的問題發生，因此需透過正規化的方式控制迴歸係數，以此降低模型變異以及樣本誤差。

2. 方法

2.1 線性迴歸(linear regression)

原始線性迴歸的目標函數為

$$\min\left\{SSE=\sum_{i=1}^{n}(y_i-\hat{y_i})^2\right\}$$

若要最佳化目標函數，資料必須符合以下幾個基本假設:

- 解釋變數與應變數之間具有線性關係
- 殘差服從常態分配
- 解釋變數之間無自相關
- 殘差變異同質性
- 觀測值個數(n)需大於解釋變數個數(p)
- 解釋變數之間不能有共線性問題

可以看到原始的線性迴歸模型有許多必須遵守的假設，因此一個替代最小平方迴歸的方法就是透過正規化迴歸(regularized regression)來控制迴歸係數。正規化迴歸模型會對迴歸係數約束，逐漸將迴歸係數壓縮至零。對迴歸係數的限制有助於降低係數的幅度和波動，降低模型的變異。

正規化迴歸的目標函數為

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

$$\text{其中 } P_\alpha(\beta)=(1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2+\alpha\|\beta\|_{l_1}=\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2+\alpha\left|\beta_j\right|\right]$$

若將$\alpha=0$時為控制係數平方，稱作 Ridge regression；若將$\alpha=1$時為控制係數絕對值，稱作 Lasso regression；若α介在 0 到 1 之間稱作 elastic net。

進行正規化迴歸模型前，會先將資料做去單位化的動作，可以得到以下兩個條件

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

有了上述兩個條件對於後續迭代式能有效縮減公式複雜度。

此時利用坐標下降法(Coordinate descent)來對目標函數做更新，該算法在每次迭代中從當前點沿一個坐標方向進行一維探索以求得一個函數的極小值。考慮$\beta_j > 0$的情況:

$$\left.\frac{\partial R_\lambda}{\partial \beta_j}\right|_{\beta=\tilde{\beta}^{(-j)}} = -\frac{1}{N}\sum_{i=1}^{N} x_{ij}\left(y_i - \tilde{\beta}_0 - x_i^T\tilde{\beta}\right) + \lambda(1-\alpha)\beta_j + \lambda\alpha \underset{set}{=} 0$$

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

經過整理後參數的迭代式為

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

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

2.2 羅吉斯迴歸(logistic regression)

有了線性的結果，另外考慮若反應變數為二元分類時該如何執行，$Y \in \{0,1\}$時，可以利用解釋變數建立其機率關係:

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

在這種分類關係上，為了使目標函數最大化，同時不使參數太多造成過度配適，採取與線性的一樣做法，加入懲罰項，使目標函數為:

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

進一步對目標函數前半段做整理

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

而要最大化這個函數時，對於運算而言過於複雜，因此利用泰勒展開式將此目標函數展至二階，做近似的結果使迭代式較簡潔得以下式子

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

其中

$$z_i = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)\left(1-\tilde{p}(x_i)\right)} \;\; ; \;\; w_i = \tilde{p}(x_i)\left(1-\tilde{p}(x_i)\right) \;\; ; \;\; C\left(\tilde{\beta}_0,\tilde{\beta}\right)^2 \;\; is \;\; constant$$

利用上述泰勒二階展開結果可以改寫目標函數為

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

此時利用坐標下降法(Coordinate descent)來對目標函數做更新，該算法在每次迭代中從當前點沿一個坐標方向進行一維探索以求得一個函數的極小值。考慮$\beta_j > 0$的情況:

$$\left.\frac{\partial R_\lambda}{\partial \beta_j}\right|_{\beta=\tilde{\beta}^{(-j)}} = \frac{1}{N}\sum_{i=1}^{N} w_i\left(z_i - \tilde{\beta}_0 - x_i^{T^{-j}}\tilde{\beta}^{(-j)} - x_{ij}\beta_j\right)(-x_{ij}) + \lambda(1-\alpha)\beta_j + \lambda\alpha \underset{set}{=} 0$$

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

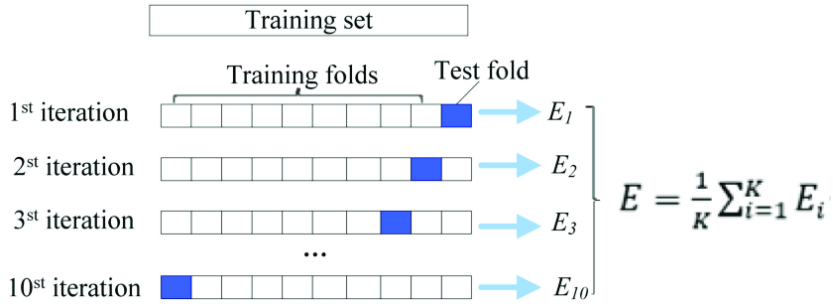
經過整理後參數的迭代式為

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

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

3. 參數調整

從前一章得到線性與非線性模型的迭代式，其中λ是需要做調整的，因此利用交叉驗證法(Cross-validation)作為調整的依據。將訓練集分割成 k 個子樣本，一個單獨的子樣本保留做為驗證模型的依據，其餘 k-1 個樣本用於訓練模型。重複此動作 k 次，直到每個子樣本皆驗證過一次，平均 k 次結果或其餘加權的結合方式，最終得到單一結果。此方法優勢在於重複運用隨機產生的子樣本進行訓練和驗證，每次結果驗證一次，可降低隨機誤差，切割成 10 份是較常用的。

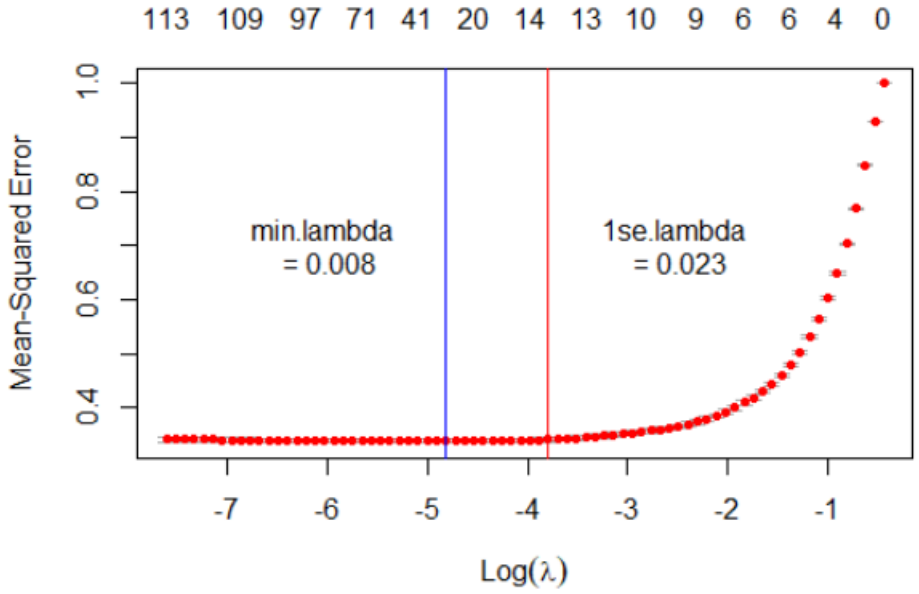


4. 實際資料

該資料共有 33334 筆，一共包含 21 個解釋變數、100 個隨機生成的變數和反應變數，其中反應變數為類別型分為 A、B 兩類，希望在使用正規化模型時，能將 100 個隨機生成的變數影響降至為 0，先將資料分割成 10000 筆的訓練資料集，最後利用剩餘的 23334 筆做為測試，計算其預測正確率(Accuracy)。

4.1 Lasso linear regression

使用此方法前，反應變數需為數值連續型，因此將資料的反應變數做轉換，轉換方式為 A 記作 1；B 記作-1。以此方式建模，利用前章所說的交叉驗證調整參數，從下圖紅色線，選擇$\lambda = 0.023$。



估計結果如下表

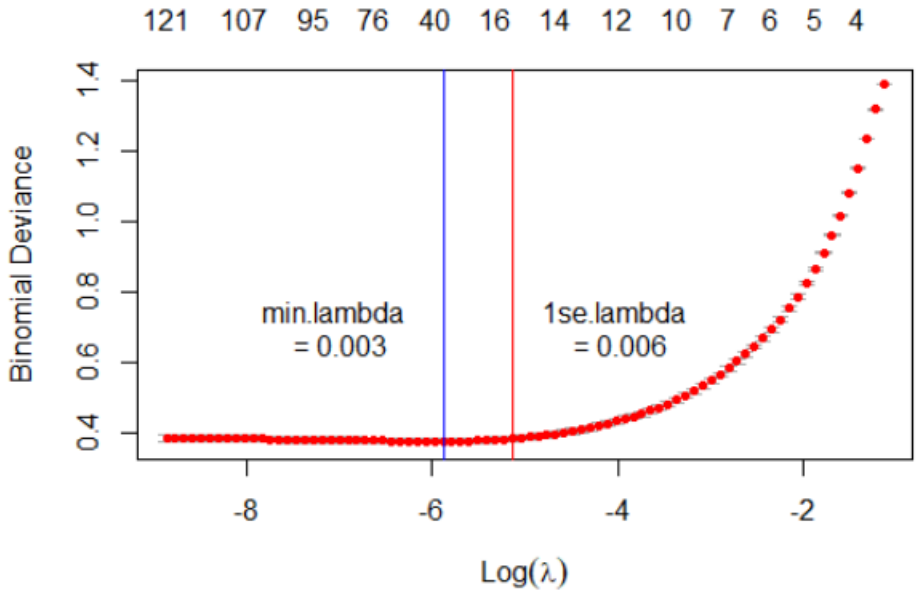
$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_4}$	$\boldsymbol{\beta_5}$	$\boldsymbol{\beta_8}$	$\boldsymbol{\beta_9}$
0.0012	0.0295	0.0417	-0.0482	-0.1355
β_{10}	β_{11}	β_{12}	β_{13}	β_{15}
-0.1556	-0.2342	-0.1115	-0.0172	0.0660
$\boldsymbol{\beta_{16}}$	β_{17}	β_{18}	β_{19}	β_{20}
0.1128	0.1410	0.891	0.0370	0.0052

此時根據此模型將 23334 的測試資料集套入，因預測出的結果會是數值型，因此將預測結果大於 0 記作 A；反之小於 0 記作 B，畫出混淆矩陣如下表，此時的預測正確率為 92.09%。此外根據估計結果可知，所生成出的 100 個噪音變數階被控制到 0 不影響模型。

混淆矩陣		實際結果	
		A	B
預測結果	A	10548	633
	B	1212	10941

4.2 Lasso logistic regression

接著使用羅吉斯迴歸，將 A 記作 1；B 記作 0，一樣使用交叉驗證的方法調整參數，從下圖紅色線選擇$\lambda = 0.006$。



估計結果如下表

$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_4}$	$\boldsymbol{\beta_5}$	$\boldsymbol{\beta_8}$	$\boldsymbol{\beta_9}$	β_{10}	β_{11}	β_{12}
-0.4118	-0.2086	-0.2441	0.3901	0.7526	0.8028	1.1859	0.6035
β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	$\boldsymbol{\beta_{18}}$	$\boldsymbol{\beta_{19}}$	$\boldsymbol{\beta_{20}}$
0.0950	-0.0190	-0.5012	-0.6512	-0.7936	-0.4935	-0.2442	-0.0704

此時根據此模型將 23334 的測試資料集套入，因預測出的結果會是數值型，因此將預測結果大於 0.5 記作 A；反之小於 0.5 記作 B，畫出混淆矩陣如下表，此時的預測正確率為 92.41%。此外根據估計結果可知，所生成出的 100 個噪音變數階被控制到 0 不影響模型。

混淆矩陣		實際結果	
		A	B
預測結果	A	10707	717
	B	1053	10857

5. 文獻回顧

1. Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software, January 2010, Volume 33, Issue 1.