Metodi per la regressione penalizzata: Group Lasso regression

Apa Marco 848154[1], Conte Enrico 852679[2], Filip Sara 852864[3]

Sommario
Nel presente documento è analizzata la Group Lasso, ovvero un'estensione della regressione Lasso a gruppi di variabili che ne permette l'inclusione o l'esclusione dal modello.
Sono successivamente approfondite delle estensioni della Group Lasso: la Sparse Group Lasso, che permette di ottenere la proprietà di *sparsity* all'interno dei gruppi, e l'Overlap Group Lasso, caso particolare in cui ogni variabile può appartenere a più di un gruppo contemporaneamente.
Infine, è affrontata una casistica, tramite l'utilizzo di R language[1], che dimostra l'utilità della Group Lasso.

Keywords
Features selection — High Dimensional Data — Lasso — Group Lasso

[1] *Universita' degli Studi di Milano Bicocca, CdLM Data Science, m.apa1@campus.unimib.it*
[2] *Universita' degli Studi di Milano Bicocca, CdLM Data Science, e.conte11@campus.unimib.it*
[3] *Universita' degli Studi di Milano Bicocca, CdLM Data Science, s.filip@campus.unimib.it*

Indice

1. Introduzione

E' noto come la regressione Lasso sia utile non solo per fare uno *shrinkage*[2] dei coefficienti, ma anche come metodo di selezione delle variabili del modello, poiché alcuni coefficienti vengono stimati esattamente pari a zero, permettendo l'esclusione di variabili non rilevanti dal modello. Nel caso in cui, considerando problemi di regressione, le covariate presentino naturalmente una struttura a gruppo, è desiderabile che tutti i coefficienti all'interno di uno stesso gruppo siano uguali a zero oppure diversi da zero, ovvero che i coefficienti siano simultaneamente tutti rilevanti oppure non rilevanti. Questo comportamento permette di escludere o di includere nel modello interi gruppi di variabili; in tal modo si evita che all'interno dello stesso gruppo siano presenti alcuni coefficienti nulli e altri non nulli.
La Group Lasso è stata molto studiata recentemente, ed ha trovato diversi ambiti di applicazione, come ad esempio in biologia per la selezione dei geni e in medicina per la previsione del peso alla nascita.

2. Group Lasso

Il principale esempio di applicazione della regressione Group Lasso si verifica quando, in una regressione lineare, tra i predittori sono presenti delle variabili categoriali (fattori); in tal caso non è opportuno usare la soluzione Lasso semplice, in quanto vengono selezionate solo le variabili dummy individuali invece che i fattori per intero. Inoltre, i risultati della regressione Lasso dipendono dalla codifica delle variabili dummy e quindi a codifiche diverse corrispondono risultati diversi [3]. La Group Lasso supera proprio questo problema attraverso un'estensione della penalità usata nella regressione Lasso.
Si considera un modello di regressione lineare con J gruppi di covariate, Z_j le covariate appartenenti al $j-esimo$ gruppo. L'obiettivo è predire y dato l'insieme delle covariate $Z_1,\dots,Z_j$. Considerando una collezione di n campioni $\{(y_i, z_{i,1}, z_{i,2}, ..., z_{i,j})\}_{i=1}^{N}$ la Group Lasso risolve il seguente problema convesso:

$$\min_{\theta_0\in\mathbb{R},\ \theta_j\in\mathbb{R}^{p_j}} \left\{ \frac{1}{2}\sum_{i=1}^{N}\left(y_i-\theta_0-\sum_{j=1}^{J} z_{ij}^T\theta_j\right)^2+\lambda\sum_{j=1}^{J}\|\theta_j\|_2 \right\} \quad (1)$$

Scritto in forma matriciale:

$$\min_{(\theta_1,\dots,\theta_j)} \left\{ \frac{1}{2}\|y - \sum_{j=1}^{J} Z_j\theta_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \right\} \tag{2}$$

In cui λ è il parametro di tuning che controlla la regolarizzazione, cioè determina di quanto i coefficienti di gruppo debbano essere ridotti verso lo zero. Come per la regressione Lasso, λ è una penalizzazione che viene inserita per controllare il trade-off tra distorsione e varianza: al crescere di λ la varianza diminuisce e la distorsione aumenta. Quando $\lambda = 0$ nessun vincolo viene imposto sulla norma euclidea e ci si riconduce così al problema dei minimi quadrati. Si nota una struttura abbastanza simile al problema di regressione Lasso (X e β nel caso di scalari, Z e θ per indicare i gruppi):

$$\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{3}$$

Osservando la penalizzazione si nota che nel caso della Lasso la penalità è in norma 1, e quindi indica la somma del valore assoluto dei coefficienti di regressione, mentre nella Group Lasso λ moltiplica la somma dei coefficienti in norma euclidea, quindi la radice quadrata della somma del quadrato dei coefficienti. Nella Group Lasso non è penalizzata la grandezza dei singoli coefficienti, ma la grandezza dei gruppi a cui appartengono.
La Group Lasso gode delle seguenti proprietà:

- In base al valore assunto da λ ($\lambda \geq 0$), tutto il vettore $\hat{\theta}_j$ (θ_j rappresenta un gruppo) è uguale a zero oppure tutte le stime dei coefficienti appartenenti al gruppo sono diverse da zero;
- Quando tutti i gruppi hanno un solo coefficiente, quindi sono formati da una sola variabile, ci si riconduce ad un normale problema di regressione Lasso.

Nella trattazione originale del problema[4], a partire da una matrice dei gruppi Z_j ortonormale, gli autori, Yuan e Lin, suggeriscono di porre $\lambda = \lambda\sqrt{p_j}$ con p_j numero di coefficienti del gruppo; in tal modo si è sicuri di applicare una penalizzazione della stessa entità sia ai gruppi grandi che a quelli piccoli. In alternativa non si introduce nessun peso, pertanto tutti i gruppi vengono penalizzati allo stesso modo (λ) e questo fa sì che sia più probabile selezionare i gruppi con numerosità maggiore.

2.1 Calcolo della Group Lasso

La funzione obiettivo della Group Lasso può essere scritta in termini matriciali:

$$\min_{(\theta_1,\dots,\theta_j)} \left\{ \frac{1}{2}\|y - \sum_{j=1}^{J} Z_j\theta_j\|_2^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \right\} \tag{4}$$

La norma euclidea è una funzione sublineare e da ciò segue che è una funzione convessa, pertanto può essere ottimizzata, ma, non essendo differenziabile nello 0, non è possibile derivare la funzione.
La funzione obiettivo 4 è convessa, ma con una componente non differenziabile, perciò la soluzione ottimale può essere trovata sfruttando le equazioni del sottogradiente, che nel caso della Group Lasso sono:

$$-Z_j^T\left(y - \sum_{l=1}^{J} Z_l\hat{\theta}_j\right) + \lambda\hat{s}_j = 0 \tag{5}$$

S_j è un elemento del subdifferenziale della norma euclidea valutata in θ_j:

$$\hat{s_j} = \begin{cases} \frac{\hat{\theta}_j}{\|\hat{\theta}_j\|_2} & \hat{\theta}_j \neq 0 \\ \in\ s:\ \|\hat{s_j}\|_2 \leq 1 & \hat{\theta}_j = 0 \end{cases}$$

La funzione obiettivo può essere minimizzata attraverso il *blockwise gradient descent*; l'algoritmo ottimizza la funzione obiettivo su un gruppo di variabili θ_j ad ogni iterazione, mantenendo fissi tutti gli altri gruppi θ_k con $k \neq j$. Applicando ciclicamente tale metodo è possibile minimizzare la funzione obiettivo. Poiché il problema è convesso e la penalità è separabile a blocchi, è garantita la convergenza ad una soluzione ottima. Fissati tutti i θ_k possiamo scrivere:

$$-Z_j^T\left(r_j - Z_j\hat{\theta}_j\right) + \lambda\hat{s}_j = 0 \tag{6}$$

con r_j residuo parziale e da cui:

$$\hat{\theta}_j = \left(Z_j^T Z_j + \frac{\lambda}{\|\hat{\theta}_j\|_2} I\right)^{-1} Z_j^T r_j \tag{7}$$

Altrimenti se $\|z_j^T r_j\|_2 < \lambda$ allora $\hat{\theta}_j = 0$.
E' possibile notare una certa somiglianza con lo stimatore ridge:

$$\hat{\beta} = \left(X^T X + \lambda I_p\right)^{-1} X^T y \tag{8}$$

I due stimatori differiscono per il fatto che nella soluzione della Group Lasso la penalità λ è divisa per una quantità pari a $\frac{1}{\|\hat{\theta}_j\|_2}$.
Se Z_j è ortonormale allora lo stimatore è:

$$\hat{\theta}_j = \left(1 - \frac{\lambda}{\|Z_j^T r_j\|_2}\right)_+ Z_j^T r_j \tag{9}$$

Un approccio alternativo consiste nell'applicazione del *proximal gradient method*; tale metodo serve proprio a minimizzare le funzioni obiettivo che sono composte da una parte convessa e differenziabile e da una parte convessa ma non differenziabile, come appunto nel caso della nostra funzione obiettivo.

Si presentano di seguito alcuni esempi di applicazione della Group Lasso.

2.2 Regressione con fattori multilivello

Si considera un problema di regressione lineare in cui si ha un predittore continuo X e un fattore G a tre livelli (g_1, g_2, g_3). Il modello lineare per la media è:

$$E(Y|X,G) = X\beta + \sum_{k=1}^{3} \theta_k \mathbb{I}_k[G] \tag{10}$$

L'equazione 10 è una regressione lineare in X con diverse intercette θ_k che dipendono dal livello di G. Usando il vettore $Z = (Z_1, Z_2, Z_3)$ composto da tre variabili dummy, uguale a $Z_k = \mathbb{I}[G]$, si può riscrivere il modello come una regressione lineare standard:

$$E(Y|X,G) = E(Y|X,Z) = X\beta + Z^T\theta \tag{11}$$

In cui $\theta = (\theta_1, \theta_2, \theta_3)$ e Z è la variabile di gruppo che include tutti i tre livelli del fattore.
Se il fattore G non ha alcun effetto sulla previsione, cioè non contribuisce a spiegare Y, allora la Group Lasso porrà l'intero vettore θ pari a 0 e verrà escluso dal modello. Se invece G ha un ruolo nella previsione di Y, ci si aspetta che la Group Lasso stimi tutti i coefficienti di θ con valori diversi da zero.
Possiamo generalizzare il modello nel caso in cui questo sia composto da diversi predittori continui e da diverse variabili di gruppo:

$$E(Y|X,G_1,..,G_J) = \beta_0 + X^T\beta + \sum_{j=1}^{J} Z_j^T\theta_j \tag{12}$$

Se non si introduce una penalizzazione in una regressione lineare che contiene fattori, bisogna preoccuparsi della distorsione; la Group Lasso supera tale problema introducendo una penalità in norma 2, che fa sì che i coefficienti in un gruppo sommino a 0.

2.3 Regressione multivariata

A volte ci si può trovare nel campo del *multitask learning*, in cui si vuole predire una variabile risposta Y multivariata dato un vettore di predittori X; ci si trova quindi nel caso in cui il problema di regressione è composto da più di una variabile dipendente e da più di una variabile esplicativa.
Date n osservazioni $\{(y_i, x_i)\}_{i=1}^{N}$, Y è la matrice della variabile risposta, X è la matrice dei predittori.
Un modello lineare può essere scritto in forma matriciale come:

$$Y = X\Theta + E \tag{13}$$

Si riconduce il problema ad una collezione di K problemi di regressione, tutti con le stesse covariate, in cui θ_k è il vettore dei coefficienti per il $k-esimo$ problema. Si può pensare di risolvere il problema ricorrendo alla regressione Lasso, ma in molte applicazioni i componenti del vettore della variabile risposta sono altamente correlati tra loro e quindi ci si aspetta che lo siano anche i vettori di regressione. E' noto che se i regressori sono molto correlati tra loro, la regressione Lasso può selezionare le variabili in modo non corretto, ovvero potrebbe selezionare delle variabili che non sono davvero rilevanti al fine della spiegazione del fenomeno. Tale comportamento è dovuto al fatto che la condizione di irrepresentabilità non è rispettata, poiché la matrice X presenta un grado di dipendenza troppo elevato.
Per affrontare questo problema si può ricorrere al Lasso adattativo oppure si possono risolvere i problemi di regressione insieme, imponendo una struttura di gruppo sui coefficienti.
Per esempio si suppone che ci sia un sottoinsieme non noto $S \subset \{1,2,..,p\}$ di variabili rilevanti per la previsione di Y e che tale sottoinsieme venga mantenuto per tutte le componenti della variabile risposta Y; si risolve il problema con la Group Lasso, definendo i p gruppi come le righe della matrice dei coefficienti θ.
Nel caso di un problema multivariato la funzione obiettivo per la Group Lasso diventa:

$$\min_{\Theta \in \mathbb{R}^{pxK}} \left\{ \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda \left(\sum_{j=1}^{p} \|\theta_j'\|_2 \right) \right\} \tag{14}$$

dove F indica la norma di *Frobenius*[5].

3. Sparse Group Lasso

La tecnica del Group Lasso, come è stato affrontato, permette di arrivare ad una formulazione sparsa della matrice dei coefficienti dei gruppi di covariate. In particolare, si ottiene che i coefficienti dei gruppi $i-esimi$ sono, alternativamente, o tutti 0 o tutti diversi da 0 a causa della non differenziabilità della $\|.\|_2$ valutata in 0. Tuttavia, a volte può essere utile arrivare ad una formulazione della matrice dei coefficienti che non sia sparsa solo *tra* i gruppi ma anche all'*interno* di essi. In termini pratici, all'interno dei gruppi con coefficienti tutti diversi da 0 viene applicato un ulteriore *shrinkage* dei coefficienti come è possibile vedere nella figura 1. In questo modo si riesce ad individuare, ad esempio, in un pathway genetico, i singoli geni che risultano importanti all'interno del gruppo che viene riconosciuto in prima istanza come rilevante.

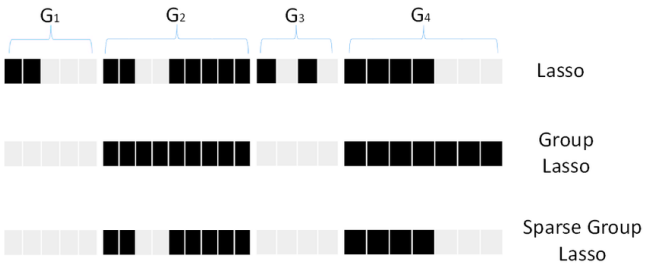

Figura 1. Esempi di gruppi delle variabili nei modelli analizzati

Nella figura 1 è possibile visualizzare quattro gruppi di variabili di esempio, in cui le colonne oscurate rappresentano le variabili scelte per il modello, mentre quelle grigie

corrispondono a quelle scartate; per ogni riga è presente la tipologia di Lasso utilizzata. Si nota come nel caso della Lasso, all'interno di uno stesso gruppo, possono essere scelte alcune variabili mentre altre no. Situazione opposta per la Group Lasso presentata, le variabili appartenenti ad uno stesso gruppo vengono tutte scelte o tutte escluse dal modello. Infine nel caso della Sparse Group Lasso si vede come all'interno dei gruppi con coefficiente diverso da 0 avvenga un ulteriore *shrinkage*.

3.1 Risoluzione del problema di minimo vincolato

Si consideri un vettore di n variabili risposta y e una matrice delle coviariate Z di dimensione (n,p) suddivisa in m sottomatrici $Z_1, Z_2, ...Z_m$ con ogni Z_l una matrice (n, p_l) dove p_l è il numero di covariate nel gruppo $l-esimo$. Il problema di minimizzazione prevede di selezionare $\hat{\theta}$ che minimizzi:

$$\frac{1}{2n}\|y-\sum_{l=1}^{m} Z_l\theta_l\|_2^2+(1-\alpha)\lambda\sum_{l=1}^{J}\|\theta_l\|_2+\alpha\lambda\|\theta\|_1 \quad (15)$$

Differentemente dalla funzione obbiettivo nel caso della Group Lasso, in questo caso il termine di penalità λ è combinazione lineare della norma 2 e della norma 1 del vettore dei coefficienti θ_l.
Come verrà esposto nella sezione successiva, questo fa si che all'interno di ogni gruppo di coefficienti avvenga un ulteriore *shrinkage* degli stessi.
Per avere più chiara la differenza tra la regione dei vincoli della Lasso, della Group Lasso e della Sparse Group Lasso si veda la figura 2.

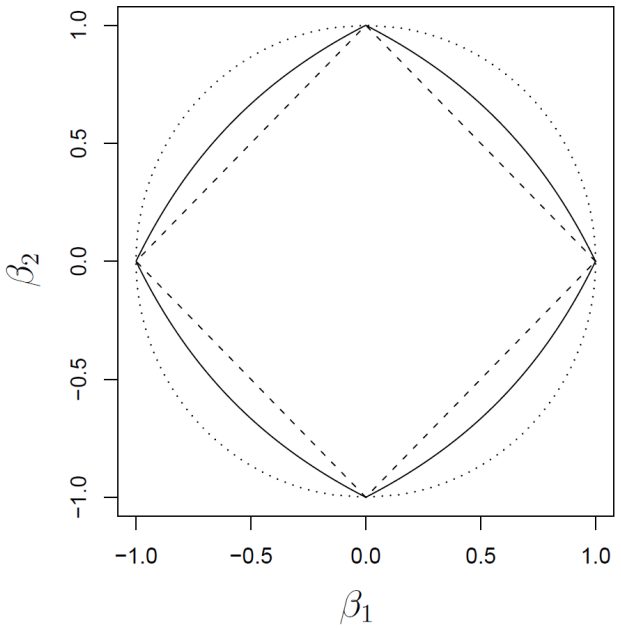

Figura 2. Forma dei vincoli dei modelli analizzati

Si noti come la diversa forma dei vincoli sia dovuta alla diversa penalizzazione imposta nel modello; dall'esterno verso l'interno $||.||_2$ per la Group Lasso (linea punteggiata), $(1-\alpha)||.||_2+\alpha||.||_1$ per la Sparse Group Lasso (linea continua) e $||.||_1$ per la Lasso (linea tratteggiata).

Dato che la funzione obbiettivo è convessa, si può sfruttare l'equazione del sottogradiente per trovare la soluzione ottima del problema.

Definito r_{-k} il valore dei residui parziali di y rispetto al gruppo k risulta:

$$r_{-k}=y-\sum_{l\neq k} Z_l\hat{\theta}_l \quad (16)$$

Per ogni gruppo k, $\hat{\theta}_k$ deve soddisfare:

$$\frac{1}{n}{Z_k}^T r_{(-k)}=(1-\alpha)\lambda u+\alpha\lambda v \quad (17)$$

Dove u e v corrispondono, rispettivamente, ai sottogradienti di $\|\hat{\theta}_k\|_2$ e $\|\hat{\theta}_k\|_1$ valutati in $\hat{\theta}_k$ tale che:

$$u=\begin{cases}\frac{\hat{\theta}_k}{\|\hat{\theta}_k\|_2} & \hat{\theta}_k\neq 0\\ \in u:\|u\|_2\leq 1 & \hat{\theta}_k=0\end{cases}$$

$$v=\begin{cases}segno(\hat{\theta}_j^k) & \hat{\theta}_j^k\neq 0\\ \in v:|v|_1\leq 1 & \hat{\theta}_j^k=0\end{cases}$$

Eseguiti i passaggi algebrici, come descritti da [6], si arriva alla conclusione che le equazioni sono soddisfatte con $\hat{\theta}_k=0$ se:

$$\|S({Z_k}^T r_{(-k)}/n,\alpha\lambda\|_2<(1-\alpha)\lambda \quad (18)$$

Con $S(*)$ operatore di soglia in funzione delle coordinate in input tale che:

$$(S(z,\alpha\lambda))_l=segno(z_j)(|z_j|-\alpha\lambda) \quad (19)$$

Mentre, per $\theta_i^k\neq 0$ si avrà che θ_i^k soddisfa:

$$\hat{\theta_i^k}=\frac{S({Z_k}^T r_{(-k)}/n,\alpha\lambda)}{{Z_l^k}^T Z_l^k/n+(1-\alpha)\lambda/\|\theta^k\|_2} \quad (20)$$

Il metodo di risoluzione rimane il *blockwise gradient descent* come nel caso della Group Lasso: per ogni gruppo k di parametri, si tengono tutti gli altri fissi e si minimizza la funzione rispetto al $k-esimo$ gruppo. Applicando iterativamente questo processo, data la convessità della funzione obbiettivo, il problema convergerà sicuramente verso la soluzione ottima.

4. Overlap Group Lasso

La versione Overlap della Group Lasso è l'estensione concettuale della versione presentata al capitolo 1, in particolare si considera la possibilità per le variabili di far parte di uno o più gruppi contemporaneamente; replicando le variabili si costruiscono i gruppi a cui successivamente si possono applicare le metodologie e le tecniche presentate per la Group Lasso.
L'appartenenza di una data variabile X_n a più di un gruppo contemporaneamente, implica che, complessivamente, le probabilità della suddetta variabile X_n di venire inclusa nel modello finale aumentino. Ciò è determinato dal fatto che il coefficiente della variabile X_n viene ora definito come la somma dei coefficienti stimati nei k gruppi nei quali la variabile X_n è presente. Ne consegue che è sufficiente che almeno uno di questi coefficienti stimati sia diverso da 0 affinché X_n venga inclusa nel modello finale. Per maggiore chiarezza si consideri il seguente esempio.

Sono presenti cinque variabili ripartite in due gruppi tali che $Z_1 = (X_1, X_2, X_3)$ e $Z_2 = (X_3, X_4, X_5)$. In questo caso la variabile X_3 è stata replicata e appartiene ad entrambi i gruppi. Si effettua successivamente il fitting dei vettori $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ e $\theta_2 = (\theta_{21}, \theta_{22}, \theta_{23})$ attraverso la formula classica della Group Lasso (1) alla quale però si modifica il termine di penalità, che viene ora definito come $\|\theta_1\|_2 + \|\theta_2\|_2$. Il coefficiente originale β_3 della variabile X_3 è definito come la somma di θ_{13} e θ_{21}. E' facile intuire come basti che uno solo tra i due coefficienti sia diverso da 0 affinché X_3 venga inclusa nel modello di regressione.

L'Overlap Group Lasso con replica delle variabili è stata la soluzione di Jacob, Obozinski and Vert (2009)[7] al problema della possibile non convergenza alla soluzione ottima della funzione nel caso di replica dei coefficienti; infatti la penalità come somma non risulta separabile. Si denota con $v_j \in \mathbb{R}^p$ il vettore composto da tutti valori nulli a meno della posizione dei membri che fanno parte del $j-esimo$ gruppo di variabili, con $V_j \in \mathbb{R}^p$ il sottospazio di tutti i possibili vettori così costruiti; considerando invece le variabili originali $X = (X_1, ..., X_p)$ il vettore dei coefficienti è dato dalla somma $\beta = \sum_{j=1}^{J} v_j$. L'Overlap Group Lasso risove il seguente problema di minimizzazione:

$$\min_{v_j \in \mathbb{V}_j,\, j=1,...,J} \left\{ \frac{1}{2} \|y - X\left(\sum_{j=1}^{J} v_j\right)\|_2^2 + \lambda \sum_{j=1}^{J} \|v_j\|_2 \right\} \quad (21)$$

Definendo una nuova funzione di penalità possiamo ricondurre il problema di ottimizzazione in termini delle variabili originali β:

$$\Omega_{\mathbb{V}}(\beta) := \inf_{v_j \in \mathbb{V}_j,\, \beta = \sum_{j=1}^{J} v_j} \sum_{j=1}^{J} \|v_j\|_2 \quad (22)$$

Come affrontato da Jacob, Obozinski and Vert (2009)[7] risolvere il problema di minimizzazione presentato (21) è equivalente a risolvere il seguente:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \Omega_{\mathbb{V}}(\beta) \right\} \quad (23)$$

Si noti che la penalizzazione $||.||_2$ per un coefficiente di una variabile è distribuita tra i gruppi a cui la variabile appartiene.

4.1 Modello gerarchico

La metodologia dell'Overlap Group Lasso può essere utilizzata per indicare una gerarchia nelle interazioni tra le variabili; questo si traduce nella possibilità di includere nel modello variabili solo se scelte insieme a quelle che descrivono un "main effect". Supponiamo di avere Z_1 e Z_2 che rappresentano le variabili dummy p_1 e p_2 con fattori rispettivamente G_1 e G_2; costruiamo la variabile $Z_{1:2} = Z_1 * Z_2$[8] ovvero il prodotto dei vettori p_1 x p_2 e come formulato da Lim and Hastie nel 2014[9] dobbiamo minimizzare la seguente funzione obiettivo:

$$\min_{\mu,\alpha,\tilde{\alpha}} \left\{ \frac{1}{2} \left\| y - \mu 1 - Z_1\alpha_1 - Z_2\alpha_2 - \begin{bmatrix} Z_1 & Z_2 & Z_{1:2} \end{bmatrix} \begin{bmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \alpha_{1:2} \end{bmatrix} \right\|_2^2 + \lambda \left(\|\alpha_1\|_2 + \|\alpha_2\|_2 + \sqrt{p_2\|\tilde{\alpha}_1\|_2^2 + p_1\|\tilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2} \right) \right\} \quad (24)$$

soggetta ai seguenti vincoli:

$$\sum_{i=1}^{p_1} \alpha_1^i = 0, \quad \sum_{j=1}^{p_2} \alpha_2^j = 0, \quad \sum_{i=1}^{p_1} \tilde{\alpha}_1^i = 0, \quad \sum_{j=1}^{p_2} \tilde{\alpha}_j^j = 0, \quad (25)$$

$$\sum_{i=1}^{p_1} \alpha_{1:2}^{ij} = 0 \; for \; fixed \; j, \quad \sum_{j=1}^{p_2} \alpha_{1:2}^{ij} = 0 \; for \; fixed \; i \quad (26)$$

I due differenti coefficienti α_j e $\tilde{\alpha}_j$ delle variabili Z_1 e Z_2 sono utilizzati per definire la penalità sovrapposta (come presentato ad inizio paragrafo), infatti il coefficiente delle due variabili sarà uguale alla somma dei rispettivi coefficienti $\theta_j = \alpha_j + \tilde{\alpha}_j$. La stima del termine di penalizzazione (24) soddisfa una forte gerarchia, infatti $\hat{\tilde{\alpha}}_1 = \hat{\tilde{\alpha}}_2 = \hat{\alpha_{1:2}} = 0$ oppure tutti i coefficienti sono diversi da 0; in definitiva o sono presenti entrambi i coefficienti relativi alle variabili che descrivono il "main effect" o quest'ultime non vengono incluse nel modello.
Si dimostra che il problema vincolato presentato si può ridurre al problema senza vincoli come segue:

$$\min_{\mu,\beta} \left\{ \frac{1}{2} \|y - \mu 1 - Z_1\beta_1 - Z_2\beta_2 - Z_{1:2}\beta_{1:2}\|_2^2 + \lambda \left(\|\beta_1\|_2 + \|\beta_2\|_2 + \|\beta_{1:2}\|_2 \right) \right\} \quad (27)$$

Quindi ogni volta che la variabile $Z_{1:2}$ è inclusa nel modello allora anche le variabili Z_1 e Z_2 che descrivono il "main effect"

saranno incluse implicitamente nel modello.
Il modello gerarchico come presentato può essere esteso naturalmente a più di una coppia di variabili.

5. Esempio di Group Lasso nel caso della regressione logistica

Come esempio di applicazione pratica della tecnica della regressione penalizzata Group Lasso, si consideri il problema della predizione del sito di splicing all'interno di una stringa di DNA. In particolare, si consideri un dataset contenente 400 osservazioni, ogni riga contiene della variabili categoriche dall'insieme di basi azotate $\{A, G, C, T\}$. Le otto colonne, invece, indicano rispettivamente le sette possibili posizioni di splicing lungo la stringa di DNA e una variabile di risposta binaria y codificata in modo che valga 0 quando la riga non corrisponde ad un sito di splicing e 1 nel caso opposto. L'obbiettivo è quello di fare predizioni riguardo i possibili siti di splicing data una particolare sequenza di $\{A, G, C, T\}$ nelle sette posizioni sopra-menzionate.
Dovendo considerare tutto l'insieme delle possibili variabili per ogni posizione, la dimensione della matrice Z dei coefficienti può diventare proibitiva. Esattamente per questo motivo si risolve il problema di minimizzazione aggiungendo una penalizzazione sulle variabili raggruppate in base alla Posizione. Così facendo, si arriva ad una formulazione sparsa della matrice dei coefficienti e il modello finale risulta più parsimonioso e interpretabile, mostrando capacità superiori di generalizzazione rispetto alla risoluzione "standard" tramite il metodo dei minimi quadrati non penalizzati. Si ottiene un modello con maggiore bias e con minore varianza delle predizioni.
L'esperimento porta alla costruzione di un modello che raggiunge il $92,5\%$ di Accuracy nel test set e un valore di Area Under the Curve (AUC) del $99,62\%$. Di contro, un modello senza termine di penalizzazione raggiunge un livello di Accuracy sullo stesso test set dell'80% e un valore di AUC di $98,46\%$.

6. Conclusioni

Si è visto che nel caso di gruppi di variabili è più conveniente usare la Group Lasso rispetto alla regressione Lasso per fare selezione di variabili, proprio perché permette di includere o escludere dal modello direttamente l'intero gruppo. Si è discusso di come la Group Lasso faccia parte dei metodi di regressione penalizzata, poiché presenta una penalizzazione λ che controlla il trade-off tra distorsione e varianza.
Inoltre sono state prese in considerazione delle estensioni della Group Lasso, ovvero la Sparse Group Lasso che risolve il problema della sparsità nei gruppi di variabili e la Overlap Group Lasso che permette alle variabili di essere presenti in più gruppi contemporaneamente e di creare gerarchie tra loro.
Complessivamente è stato visto come queste tecniche di regressione penalizzata permettano di ottenere dei modelli che generalizzano meglio in virtù del fatto che viene ceduta una parte della varianza a favore della distorsione.

Riferimenti bibliografici

[1] Wikipedia contributors. R (programming language) — Wikipedia, the free encyclopedia, 2019. [Online; accessed 20-June-2019].

[2] Wikipedia contributors. Shrinkage (statistics) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Shrinkage_(statistics)&oldid=981225319, 2020. [Online; accessed 8-December-2020].

[3] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso, 2010.

[4] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[5] Eric W. Weisstein. Frobenius norm. — from MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/FrobeniusNorm.html, 2020.

[6] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[7] L. Jacob, G. Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *ICML '09*, 2009.

[8] Wikipedia contributors. Hodge star operator — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Hodge_star_operator&oldid=986539900, 2020. [Online; accessed 9-December-2020].

[9] Hastie T Lim M. Learning interactions via hierarchical group-lasso regularization. pages 627–654, 2014.