Seminarium magisterskie 2

LASSO oraz powiązane algorytmy ścieżkowe

Magdalena Trafidło

13 kwietnia 2023

Przypadek wielowynikowy - metody regularyzacji i wyboru

Niech Y - macierz wynikowa z K kolumnami.

$$Y_k = f(x) + \epsilon_k,$$

$$Y_\ell = f(x) + \epsilon_\ell$$

Analiza korelacji kanonicznych (CCA)

CCA znajduje sekwencję X_{v_m} - nieskorelowanych liniowych kombinacji x_j oraz odpowiadające im Y_{u_m} - nieskorelowane liniowe kombinacje y_k, tak aby

$$Corr^2(Y_{u_m}, X_{v_m})$$

było jak największe.
Do rozwiązania powyższego problemu stosuje się SVD na macierzy $Y^T X/N$, gdzie X, Y - scentrowane.

Reduced-rank regression

$$\hat{B}^{rr} = \underset{rank(B)=m}{\arg\min} \sum_{i=1}^{N} (y_i - B^T x_i)^T \Sigma^{-1} (y_i - B^T x_i), \Sigma = Cov(\epsilon).$$

Jeśli zastąpimy Σ przez estymator $Y^T Y/N$, dostaniemy rozwiązanie wyznaczone przez CCA:

$$\hat{B}^{rr}(M) = (X^T X)^{-1} X^T (Y U_m) U_m^-,$$

gdzie U_m jest podmacierzą rozmiaru $K \times m$ macierzy U.

Incremental Forward Stagewise Regression

Incremental Forward Stagewise Regression - przyrostowa regresja stopniowa do przodu

Algorytm FS_ϵ

1. Przyjmij na początku, że residua r są równe y, a $\beta_0, \ldots, \beta_p = 0$. Dodatkowo, niech predyktory będą ustandaryzowane (średnia zero, norma równa 1).
2. Znajdź predyktor x_j najbardziej skorelowany z r.
3. Przypisz $\beta_j \leftarrow \beta_j + \delta_j$, $r_j \leftarrow r_j - \delta_j x_j$, gdzie $\delta_j = \epsilon \cdot \text{sign}|\langle x_j, r\rangle|$, a $\epsilon > 0$ to ustalony krok.
4. Powtarzaj 2-3 do momentu, w którym residua będą nieskorelowane ze wszystkimi predyktorami.

Zauważmy, że jeśli $\delta_j = \langle x_j, r\rangle$, to dostaniemy algorytm *Forward Stagewise*.

Infinitesimal Forward Stagewise Regression ($\epsilon \longrightarrow 0$)

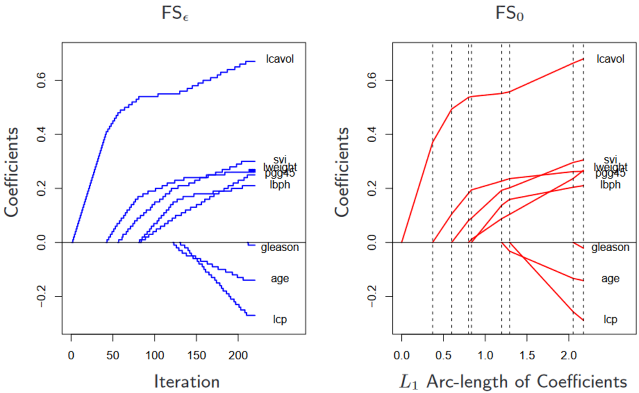
Algorytm LAR - modyfikacja FS_0

1. Przyjmij na początku, że $r = y - \bar{y}$, a $\beta_0, \ldots, \beta_p = 0$. Dodatkowo, niech predyktory będą ustandaryzowane (średnia zero, norma równa 1).
2. Znajdź predyktor x_j najbardziej skorelowany z r.
3. Przesuwaj β_j od zera w kierunku współczynnika $\langle x_j, r \rangle$, do momentu, w którym jakiś predyktor x_k będzie tak samo skorelowany z obecnymi residuami jak x_j.
4. Znajdź nowy kierunek rozwiązując następujący problem:

   $$\min_b \|r - X_A b\|_2^2 \text{ z } b_j s_j \geqslant 0, j \in A,$$

   a s_j to znak $\langle x_j, r \rangle$.
5. Kontynuuj tak długo, aż zostaną wykorzystane wszystkie predyktory. Po $\min(N-1, p)$ krokach koniec.

Przykład - rak prostaty, $\epsilon = 0.01$



Algorytmy ścieżkowe kawałkami liniowe

$\hat{\beta}(\lambda)$ - ścieżka:

$$\hat{\beta}(\lambda) = \arg\min_{\beta}[R(\beta) + \lambda J(\beta)], \text{ gdzie}$$

$$R(\beta) = \sum_{i=1}^{N} L\left(y_i, \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right),$$

L - funkcja straty, J - funkcja kary - wypukłe. Jeśli:

1. R jest kwadratową lub miejscami kwadratową funkcją β,
2. J jest miejscami liniowa w β

to rozwiązanie $\hat{\beta}(\lambda)$ będzie kawałkami liniowe.

Selektor Dantziga

Kryterium *Dantzig Selector (DS)*:

$$\min_{\beta} \|\beta\|_1 \text{ przy warunku } \|X^T(y - X\beta)\|_\infty \leqslant s,$$

lub równoważnie:

$$\min_{\beta} \|X^T(y - X\beta)\|_\infty \text{ przy warunku } \|\beta\|_1 \leqslant t,$$

$\|\cdot\|_\infty$ - norma maksimum, czyli maksimum z wartości bezwzględnych elementów wektora.
Jeśli $t \to \infty$ i $N < p$, rozwiązanie zbiega do rozwiązania uzyskanego przy pomocy metody najmniejszych kwadratów, jeśli $N \geqslant p$ rozwiązanie zbiega do rozwiązania uzyskanego przy pomocy metody najmniejszych kwadratów z minimum normy L_1.

Grupowe Lasso

L - liczba grup, na które podzielono p predyktorów,
p_ℓ - liczba predyktorów w grupie ℓ,
X_ℓ - macierz predyktorów odpowiadająca grupie ℓ,
β_ℓ - wektor odpowiadający X_ℓ.

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^p} \left(\|y - \beta_0 1 + \sum_{\ell=1}^{L} X_\ell \beta_\ell\|_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} \|\beta_\ell\|_2 \right).$$

Smoothly clipped absolute deviation (SCAD)

Zmieniamy karę z $\lambda|\beta|_1$ na $J_a(\beta, \lambda)$, czyli szukamy $\hat{\beta}$:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + J_a(\beta, \lambda) \right\},$$

gdzie

$$\frac{dJ_a(\beta, \lambda)}{d\beta} = \lambda \cdot \mathrm{sign}(\beta)[I(|\beta| \leqslant \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)], a \geqslant 2.$$

Pathwise Coordinate Optimization

Zakładamy, że predyktory są ustandaryzowane (ze średnią 0 i normą 1). Niech $\tilde{\beta}_k(\lambda)$ oznacza obecny estymator β_k z parametrem λ. Wtedy

$$R(\tilde{\beta}(\lambda), \beta_j) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k(\lambda) - x_{ij}\beta_j\right)^2 + \lambda\sum_{k\neq j}|\tilde{\beta}_k(\lambda)| + \lambda\beta_j,$$

co daje

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

$$S(t,\lambda) = \operatorname{sign}(t)(|t| - \lambda)_+.$$