

Linear Regression: Algorithms and Instabilities

Monday, April 10, 2017 9:32 AM

Goals of Lecture 3

1. Motivate extensions of OLS
2. Matrix decompositions
3. Ridge regression
4. Subset selection : greedy methods

Recall linear regression

$$\hat{y} = X \hat{\beta} \quad \hat{\beta} = (X^T X)^{-1} X^T y$$

$n \quad n \times p \quad p$

$$\hat{y} = \underbrace{X(X^T X)^{-1} X^T}_H y$$

H : hat matrix

$$\text{Fit: solve } X^T X \hat{\beta} = X^T y$$

Algorithm 3.1 Regression by Successive Orthogonalization.

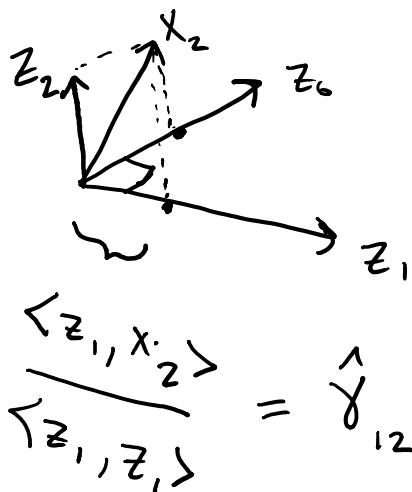
1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.

ESL pg. 54

2. For $j = 1, 2, \dots, p$

Regress \mathbf{x}_j on $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \langle \mathbf{z}_\ell, \mathbf{x}_j \rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle$, $\ell = 0, \dots, j-1$ and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.

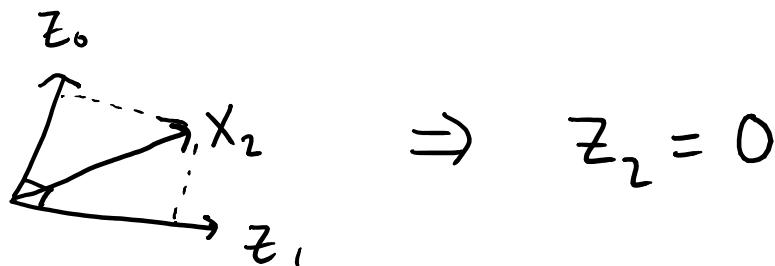
3. Regress \mathbf{y} on the residual \mathbf{z}_p to give the estimate $\hat{\beta}_p$.



\mathbf{x}_j - n dim column of \mathbf{X}

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}$$

$\swarrow \overline{\dots} \searrow$ Instable if $\langle \mathbf{z}_p, \mathbf{z}_p \rangle$ is small $\swarrow \overline{\dots} \searrow$
 Impossible if $\langle \mathbf{z}_p, \mathbf{z}_p \rangle = 0$



linear dependence (always happens if $p > n$)

Matrix Decompositions and Linear Regression

Monday, April 10, 2017 10:33 AM

Gram-Schmidt $\rightarrow X = Z \Gamma$
 orthogonal $\xrightarrow{\text{orthogonal}}$ R upper triangular

$$[x_0, x_1, \dots, x_p] = [z_0, z_1, \dots, z_p] \begin{bmatrix} 1 & \hat{y}_{01} & \hat{y}_{02} & \dots \\ 0 & 1 & \hat{y}_{12} & \dots \\ \vdots & 0 & 1 & \ddots \\ 0 & 0 & 0 & \ddots \end{bmatrix}$$

$$z_j = x_j - \sum_{k=0}^{j-1} \hat{y}_{kj} z_k$$

$$x_j = z_j + \sum_{k=0}^{j-1} \hat{y}_{kj} z_k$$

QR-decomposition D-diagonal w/ $D_{jj} = \|z_j\|$

$$X = \underbrace{Z}_Q \underbrace{D^{-1}}_R \Gamma = Q R \xrightarrow{\text{orthonormal}} \text{upper triangular}$$

$$\text{Orthogonal: } Z^T Z = \begin{bmatrix} \|z_0\|^2 & 0 \\ 0 & \ddots & \vdots \\ 0 & \cdots & \|z_p\|^2 \end{bmatrix}$$

$$\text{Orthonormal: } Q^T Q = I = \begin{bmatrix} 1 & 0 & \dots \\ 0 & 1 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

OLS

$$\hat{\beta} = (X^T X)^{-1} X^T y = (R^T Q^T Q R)^{-1} R^T Q^T y$$

$$= (R^T R)^{-1} R^T Q^T y = R^{-1} Q^T y$$

$$\hat{y} = X\hat{\beta} = QR\hat{\beta} = QRR^{-1}Q^T y = QQ^T y$$

Singular Value Decomposition ($N > p$)

Let X be centered ($x_i \leftarrow x_i - \bar{x}_i$) then

$$X = UDV^T$$

with U - orthonormal $N \times p$

V - orthonormal $p \times p$

D - diagonal $p \times p$

▷ $X^T X = (V D U^T)(U D V^T) = V D^2 V^T$

(spectral decomposition)

▷ $(X^T X)^{-1} = V D^{-2} V^T$ w/ $(D^{-2})_{jj} = D_{jj}^{-2}$

$(X^T X)^{-1} X^T X = V D^{-2} V^T V D^2 V^T = V V^T (= I_{\text{indep}})$

▷ $X\hat{\beta} = X(X^T X)^{-1} X^T y$

$$= UDV^T(VD^{-2}V^T)VDU^Ty$$

$$= UU^Ty$$

▷ more computationally intensive than QR!

▷ $\hat{\beta} = VD^{-2}V^TVDU^T y = VD^{-1}U^Ty$

$$\triangleright \hat{\beta} = V D^{-2} V^T V D U^T y = V D^{-1} U^T y$$

$\hat{\beta}$ is unstable if eigenvalues D_{ii} are small

Ridge Regression

Monday, April 10, 2017 11:59 AM

$\hat{\beta}$ if X is nearly singular then $\hat{\beta}$ is unstable.

Ridge regression

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

complexity parameter

Center X : $x_{ij} \leftarrow x_{ij} - \bar{x}_j$

$$\text{then } \hat{\beta}_0^{\text{ridge}} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Each β_j is treated the same
Normalize:

$$x_{ij} \leftarrow x_{ij} / \|x_j\|$$

$$\begin{aligned} \text{Ridge objective: } & (y - X\beta)^T(y - X\beta) + \lambda \beta^T \beta \\ & \propto -2y^T X\beta + \beta^T X^T X \beta + \lambda \beta^T \beta \\ & \stackrel{\partial \beta}{\rightarrow} 2(X^T X + \lambda I)\beta - 2X^T y \end{aligned}$$

Ridge normal equations:

$$(X^T X + \lambda I)\hat{\beta}^r = X^T y$$

$$\text{SVD: } X^T X + \lambda I = V(D^2 + \lambda I)V^T$$

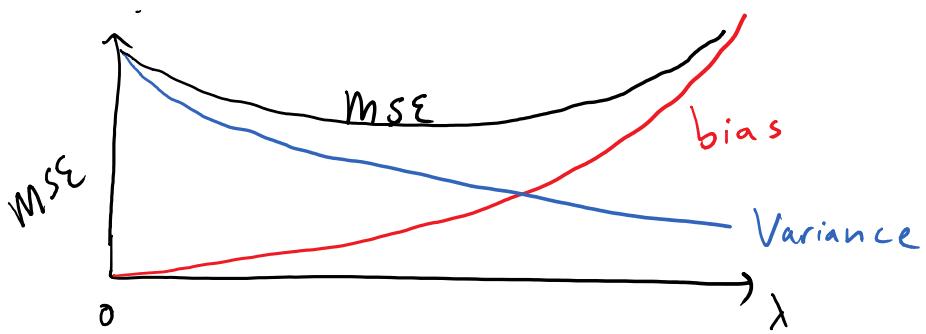
$$\text{Sln: } \hat{\beta}^r = V(D^2 + \lambda I)^{-1} D U^T y = (X^T X + \lambda I)^{-1} X^T y$$

$\hat{\beta}$ large stabilizes $\hat{\beta}^r$ (lower Variance)

$$y_i = x_i^\top \beta + \varepsilon_i \quad \mathbb{E}[\varepsilon_i | x_i] = 0 \rightarrow y = X\beta + \varepsilon$$

$$\begin{aligned}\mathbb{E}[\hat{\beta}|X] &= \mathbb{E}[(X^\top X + \lambda I)^{-1} X^\top y | X] \\ &= \mathbb{E}[(X^\top X + \lambda I)^{-1} X^\top X \beta + (X^\top X + \lambda I)^{-1} X^\top \varepsilon | X] \\ &= (X^\top X + \lambda I)^{-1} X^\top X \beta\end{aligned}$$

(X is non-singular, $\lambda = 0 \Rightarrow \mathbb{E}[\hat{\beta}|X] = \beta$ - OLS)



Subset Selection

Monday, April 10, 2017 3:12 PM

Two motivations for subset selection (selecting variables)

1. Fewer variables can lead to lower risk
(measured by holdout risk)

2. We want to discover subset of variables with large effects.

A model for sparsity: $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ (linear model)

def Support of $\boldsymbol{\beta}$, $\text{supp}(\boldsymbol{\beta})$, is $\{j : \beta_j \neq 0\}$.

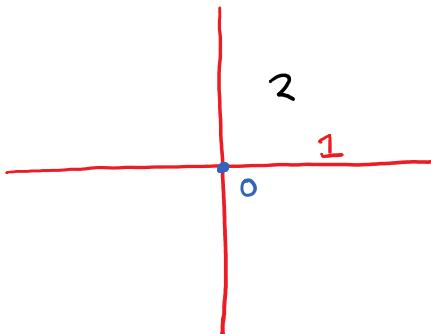
Computational Challenges

- Select $S \subseteq \{1, \dots, p\}$ s.t. $|S|=k$ that minimizes $\min_{\boldsymbol{\beta}_S} \sum_{i=1}^n (y_i - \sum_{j \in S} x_{ij} \beta_j)^2$ OLS(\mathbf{x}_S)

NP-hard

def l_0 -norm: $\|\boldsymbol{\beta}\|_0 = |\text{supp}(\boldsymbol{\beta})|$

Subset selection: $\min_{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|_0 \leq k} \|y - \mathbf{x} \boldsymbol{\beta}\|_2^2$



↳ alas, $\|\boldsymbol{\beta}\|_0$ is discontinuous and non-convex!

Greedy methods

Basic idea: at each step, choose action that improves the empirical risk

Forward Stepwise: X is standardized, $x_0 = 1$

Let $S_0 = \{0\}$, $[p] = \{1, \dots, p\}$

For $k=1, \dots, p$

For j in $[p] \setminus S_{k-1}$

$$R_j := \min_{\substack{\text{supp}(\beta) = S_{k-1} \cup \{j\}}} \|y - X\beta\|_2^2 \quad (\star)$$

$$j_k = \operatorname{argmin}_{j \in [p] \setminus S_{k-1}} R_j$$

$$S_k \leftarrow S_{k-1} \cup \{j_k\}$$

Backwards Stepwise: Start with the full model ($n < p$)
remove variables

Forward Stagewise:

Replace (\star) with $r_{k-1} = y - X\hat{\beta}_{k-1}$ (last fit residuals)

$$\min_{\beta_j \in \mathbb{R}} \|r_{k-1} - \beta_j x_j\|_2^2 \quad \left(\hat{\beta}_j = \frac{r_{k-1}^\top x_j}{\|r_{k-1}\| \|x_j\|} \right)$$