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

$$\underset{n}{\hat{y}} = \underset{n \times p}{X} \underset{p}{\hat{\beta}} \qquad \hat{\beta} = (X^T X)^{-1} X^T y$$

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

H : hat matrix

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

Algorithm 3.1 *Regression by Successive Orthogonalization.*

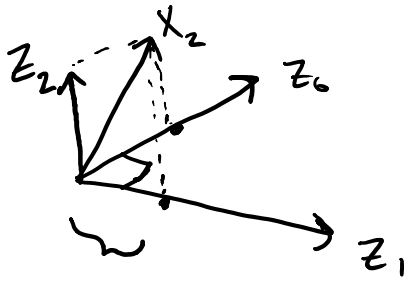
1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.

ESL pg. 54

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

   Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, , \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \langle \mathbf{z}_\ell, \mathbf{x}_j \rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle$, $\ell = 0, \ldots, j-1$ 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 $\hat{\beta}_p$.



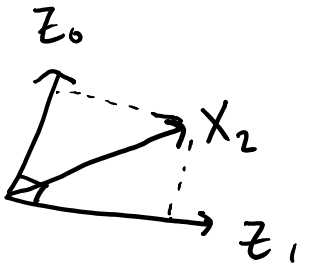
x_j – n dim column of X

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

$$\frac{\langle z_1, x_2 \rangle}{\langle z_1, z_1 \rangle} = \hat{\gamma}_{12}$$

!!! Instable if $\langle z_p, z_p \rangle$ is small !!!

Impossible if $\langle z_p, z_p \rangle = 0$



$\Rightarrow \quad z_2 = 0$

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

Matrix Decompositions and Linear Regression

Monday, April 10, 2017 10:33 AM

Gram-Schmidt $\rightarrow$ $X = Z\Gamma$

Γ: upper triangular; Z: orthogonal

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

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

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

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

$$X = \underbrace{ZD^{-1}}_{Q} \underbrace{D\Gamma}_{R} = QR$$

Q: orthonormal; R: upper triangular

orthogonal: $Z^T Z = \begin{bmatrix} \|z_0\|^2 & & 0 \\ & \ddots & \\ 0 & & \|z_p\| \end{bmatrix}$

orthonormal: $Q^T Q = I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

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 ($x_j \leftarrow x_j - \overline{x_j}$) then

$$X = UDV^T$$

with U – orthogonal $N \times p$

V – orthogonal $p \times p$

D – diagonal $p \times p$

- $X^TX = (VDU^T)(UDV^T) = VD^2V^T$

  (spectral decomposition)

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

  $(X^TX)^{-1}X^TX = VD^{-2}V^TVD^2V^T = VV^T$ ($= I$ lin indep)

- $X\hat{\beta} = X(X^TX)^{-1}X^Ty$

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

  $= UU^Ty$

- more computationally intensive than QR!

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

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

Ridge Regression

Monday, April 10, 2017 11:59 AM

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

Ridge regression

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

λ: complexity parameter

Center X: $x_{ij} \leftarrow x_{ij} - \overline{x_j}$

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

Each β_j is treated the same

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

Ridge objective: $(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$

$$\propto -2y^TX\beta + \beta^TX^TX\beta + \lambda\beta^T\beta$$

$$\xrightarrow{\partial\beta} 2(X^TX + \lambda I)\beta - 2X^Ty$$

Ridge normal equations:

$$(X^TX + \lambda I)\hat{\beta}^r = X^Ty$$

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

Soln: $\hat{\beta}^r = V(D^2 + \lambda I)^{-1}U^Ty = (X^TX + \lambda I)^{-1}X^Ty$

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

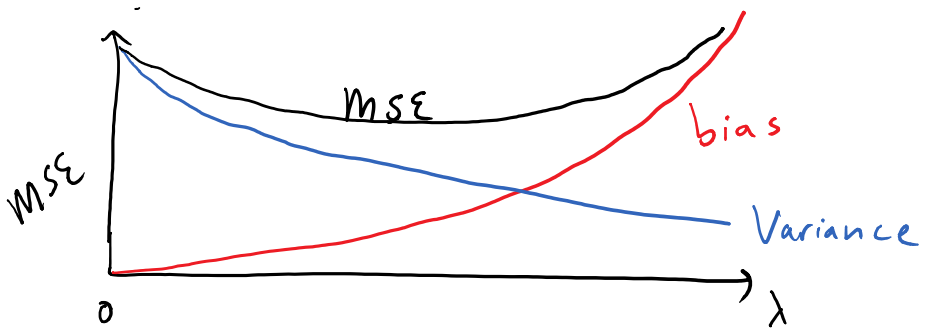
$y_i = x_i^T \beta + \varepsilon_i \quad \mathbb{E}[\varepsilon_i | x_i] = 0 \rightarrow y = X\beta + \varepsilon$

$$\mathbb{E}[\hat{\beta}^r | X] = \mathbb{E}[(X^TX + \lambda I)^{-1} X^T y | X]$$

$$= \mathbb{E}[(X^TX + \lambda I)^{-1} X^TX\beta + (X^TX + \lambda I)^{-1} X^T\varepsilon | X]$$

$$= (X^TX + \lambda I)^{-1} X^TX\beta$$

(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: $y_i = x_i^T \beta + \varepsilon_i$ (linear model)

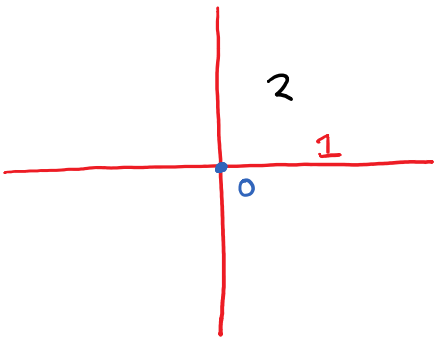
<u>def</u> Support of β, $\text{supp}(\beta)$, is $\{j : \beta_j \neq 0\}$.

Computational Challenges:

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

NP-hard

<u>def</u> ℓ_0-norm: $\|\beta\|_0 = |\text{supp}(\beta)|$



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

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

Greedy methods

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

Forward Stepwise: X is standardized, $X_0 = \mathbb{1}$

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

For $k = 1, \dots, p$

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

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

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

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

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

Forward <u>Stagewise</u>:

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

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