End-of-Semester Rambling #1

Ridge Regression

Ridge Regression

Numeric Hack Instead of

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

use

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Question Clearly makes sense when $p > n$, in which case $\boldsymbol{X}^\top \boldsymbol{X}$ is not even invertible, but beneficial sometimes even when $p < n$. Why?

Remark Really only makes sense if data are centered and scaled, i.e., $\mathbf{1}^\top \boldsymbol{y} = 0, \|\boldsymbol{y}\| = 1$ and likewise for columns of $\boldsymbol{X}$. (Why?)

Bias-Variance Analysis

Bias

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad [\text{unbiased}], \qquad \mathbb{E}(\widehat{\boldsymbol{\beta}}_\lambda) \neq \boldsymbol{\beta} \quad [\text{biased}]$$

Variance

$$\begin{aligned} \mathbb{V}\text{ar}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top [\mathbb{V}\text{ar}(\boldsymbol{y})] \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \end{aligned}$$

$$\begin{aligned} \mathbb{V}\text{ar}(\widehat{\boldsymbol{\beta}}_\lambda) &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top [\mathbb{V}\text{ar}(\boldsymbol{y})] \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \\ &= \sigma^2 [(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1}] \end{aligned}$$

Question Clearly, $\mathbb{B}\text{ias}(\widehat{\boldsymbol{\beta}}_\lambda) \geq \mathbb{B}\text{ias}(\widehat{\boldsymbol{\beta}})$, but $\mathbb{V}\text{ar}(\widehat{\boldsymbol{\beta}}_\lambda) \gtrless \mathbb{V}\text{ar}(\widehat{\boldsymbol{\beta}})$?

Singular Value Decomposition (SVD)

SVD Assume $\boldsymbol{X}$ is $n \times p$ with $n \geq p$. There exists decomposition

$$\boldsymbol{X} = \underset{n \times p}{\boldsymbol{U}} \overset{p \times p}{\downarrow} \boldsymbol{D} \underset{p \times p}{\uparrow} \boldsymbol{V}^{\top},$$

where $\boldsymbol{U}$, $\boldsymbol{V}$ orthonormal; $\boldsymbol{D}$ diagonal with all $\boldsymbol{D}_{ii} \geq 0$.

Remark

(a) Assumption $n \geq p$ not restrictive — if $n \leq p$, simply apply SVD to $\boldsymbol{X}^{\top}$ and "transpose back".

(b) Probably the most important/useful result in matrix algebra.

Bias-Variance Analysis

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \quad \Rightarrow \quad \boldsymbol{X}^\top\boldsymbol{X} = (\boldsymbol{V}\boldsymbol{D}\underbrace{\boldsymbol{U}^\top)(\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{D}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top$$

$$\Rightarrow \quad (\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\top$$

$$\Rightarrow \quad (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} = (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + \lambda\boldsymbol{V}\boldsymbol{V}^\top)^{-1}$$

$$= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}^\top$$

$$\left.\begin{aligned} \mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) &= \sigma^2\boldsymbol{V}\,\mathrm{diag}\left[\frac{1}{\boldsymbol{D}_{ii}^2}\right]\boldsymbol{V}^\top \\ \mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}_\lambda) &= \sigma^2\boldsymbol{V}\,\mathrm{diag}\left[\frac{\boldsymbol{D}_{ii}^2}{(\boldsymbol{D}_{ii}^2+\lambda)^2}\right]\boldsymbol{V}^\top \end{aligned}\right\} \frac{\boldsymbol{D}_{ii}^2}{(\boldsymbol{D}_{ii}^2+\lambda)^2} \le \frac{1}{\boldsymbol{D}_{ii}^2}$$

$$\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}_\lambda) \le \mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}})$$

ridge regression most useful in "high-variance scenarios"

Multicollinearity

If there exists $\boldsymbol{\ell} \in \mathbb{R}^p$ such that $\|\boldsymbol{\ell}\| = 1$ and $\|\boldsymbol{X}\boldsymbol{\ell}\| = \varepsilon$, then

$$\|\boldsymbol{X}\boldsymbol{\ell}\|^2 = \boldsymbol{\ell}^\top \underbrace{\boldsymbol{X}^\top \boldsymbol{X}}_{\boldsymbol{S}} \boldsymbol{\ell} = \boldsymbol{\ell}^\top \boldsymbol{S}\boldsymbol{\ell} = \|\boldsymbol{S}^{1/2}\boldsymbol{\ell}\|^2 = \varepsilon^2$$

and

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\boldsymbol{\ell}^\top \widehat{\boldsymbol{\beta}}) = \boldsymbol{\ell}^\top \mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}})\boldsymbol{\ell} &= \sigma^2 \boldsymbol{\ell}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{\ell} \\ &= \sigma^2 \boldsymbol{\ell}^\top \boldsymbol{S}^{-1} \boldsymbol{\ell} = \sigma^2 \|\boldsymbol{S}^{-1/2}\boldsymbol{\ell}\|^2.\end{aligned}$$

But the Cauchy-Schwarz inequality implies

$$\|\boldsymbol{S}^{1/2}\boldsymbol{\ell}\|^2 \|\boldsymbol{S}^{-1/2}\boldsymbol{\ell}\|^2 \geq (\boldsymbol{\ell}^\top \boldsymbol{S}^{1/2} \boldsymbol{S}^{-1/2} \boldsymbol{\ell})^2 = \|\boldsymbol{\ell}\|^4 = 1,$$

so

$$\mathbb{V}\mathrm{ar}(\boldsymbol{\ell}^\top \widehat{\boldsymbol{\beta}}) \geq \frac{\sigma^2}{\varepsilon^2}.$$

End-of-Semester Rambling #2

Generalized Linear Models

Generalized Linear Models (GLM)

Exponential Family

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - \gamma(\theta)}{\phi} + c(y;\phi)\right]$$

θ = natural parameter, ϕ = scale parameter

Canonical GLM

$y \sim$ exponential family with natural parameter $\theta = \boldsymbol{x}^\top \boldsymbol{\beta}$

y: response; $\boldsymbol{x}$: covariates

Examples

$y \in \mathbb{R} \Rightarrow y \sim \mathrm{N}(\mu, \sigma^2)$; $y \in \{0,1\} \Rightarrow y \sim \mathrm{Bernoulli}(p)$

Normal

$$\boxed{f(y;\theta,\phi) = \exp\left[\frac{y\theta - \gamma(\theta)}{\phi} + c(y;\phi)\right]}$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

$$= \exp\left[\frac{y\mu - \overbrace{\mu^2/2}^{\gamma(\cdot)}}{\sigma^2} + \overbrace{\log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{y^2/2}{\sigma^2}}^{c(y,\cdot)}\right]$$

θ	ϕ	$\gamma(\cdot)$	$c(y,\cdot)$
μ	σ^2	$\mu^2/2$	$\log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{y^2/2}{\sigma^2}$

Bernoulli

$$\boxed{f(y;\theta,\phi) = \exp\left[\frac{y\theta - \gamma(\theta)}{\phi} + c(y;\phi)\right]}$$

$$f(y) = p^y(1-p)^{1-y} = \exp\left[y\log(p) + (1-y)\log(1-p)\right]$$

$$= \exp\left[y\log\frac{p}{1-p} + \underbrace{\log(1-p)}_{-\gamma(\cdot)}\right]$$

θ	ϕ	$\gamma(\cdot)$	$c(y,\cdot)$
$\log[p/(1-p)]$	1	$-\log(1-p)$	0

$$\theta = \log\frac{p}{1-p} \quad \Rightarrow \quad p = \frac{e^\theta}{1+e^\theta} \quad \Rightarrow \quad \gamma(\theta) = \log(1+e^\theta)$$

Logit vs Sigmoid Transforms

Logit

$$\theta = \log \frac{p}{1-p} = \begin{cases} +\infty, & p = 1 \\ 0, & p = 1/2 \\ -\infty, & p = 0 \end{cases}$$

Sigmoid

$$p = \frac{e^\theta}{1+e^\theta} = \begin{cases} 1, & \theta \to +\infty \\ 1/2, & \theta = 0 \\ 0, & \theta \to -\infty \end{cases}$$

Canonical GLMs

- $y_1, y_2, ..., y_n \in \mathbb{R} \Rightarrow$ "ordinary" regression

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

where

$$\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

- $y_1, y_2, ..., y_n \in \{0, 1\} \Rightarrow$ logistic regression

$$y_i \sim \text{Bernoulli}(p_i)$$

where

$$\log \frac{p_i}{1 - p_i} = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \text{or} \quad p_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}$$

Logistic Regression

- log-likelihood:

$$\sum_{i=1}^{n} y_i \log \frac{p_i}{1-p_i} + \log(1-p_i)$$

$$\log \frac{p_i}{1-p_i} = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \Rightarrow \quad \Downarrow \quad \Leftarrow \quad p_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}$$

$$\ell(\boldsymbol{\beta}; y_i, \boldsymbol{x}_i) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})$$

- MLE:

$$\max_{\boldsymbol{\beta}} \quad \ell(\boldsymbol{\beta}; y_i, \boldsymbol{x}_i)$$

by Newton-Raphson

Newton-Raphson

$$\ell(\boldsymbol{\beta}; y_i, \boldsymbol{x}_i) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})$$

$$\Downarrow$$

$$\ell'(\boldsymbol{\beta}) = \quad \ldots \quad = \sum_{i=1}^{n} (y_i - p_i)\boldsymbol{x}_i$$

$$= \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} y_1 - p_1 \\ \vdots \\ y_n - p_n \end{bmatrix} = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p})$$

Newton-Raphson

$$
\begin{aligned}
& \ell''(\boldsymbol{\beta}) \\
= & \ \vdots \\
= & -\begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} p_1(1-p_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_n(1-p_n) \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \\
= & -\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}
\end{aligned}
$$

Iteratively Reweighted LSQ

$$
\begin{aligned}
\boldsymbol{\beta}_{new} &= \boldsymbol{\beta}_{old} - \left[\ell''(\boldsymbol{\beta}_{old})\right]^{-1}\left[\ell'(\boldsymbol{\beta}_{old})\right] \\
&= \boldsymbol{\beta}_{old} + \left[\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right]^{-1}\left[\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{p})\right] \\
&= \underbrace{\left[\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}}_{\text{weighted LSQ operator}} \underbrace{\left[\boldsymbol{X}\boldsymbol{\beta}_{old} + \boldsymbol{W}^{-1}(\boldsymbol{y}-\boldsymbol{p})\right]}_{\boldsymbol{z};\ \text{working response}}
\end{aligned}
$$

(notice that both $\boldsymbol{W}$ and $\boldsymbol{p}$ depend on $\boldsymbol{\beta}_{old}$ as well)

$\boldsymbol{W}_{ii} \equiv p_i(1-p_i)$ max at $p_i = 1/2$ and min at $p_i = 0$ or $p_i = 1$

$\Downarrow$

estimates influenced mostly by points near decision boundary