LECTURE TWELVE

Model:

$$y_t = \beta_0 + \beta_1 (t-1) + \beta_2 \text{ReLU}(t-2) + \cdots + \beta_{n-1} \text{ReLU}(t-(n-1)) + \varepsilon_t$$

$t = 1, \ldots, n$

→ high-dimension linear regression

$$\boxed{y = X\beta + \varepsilon}$$

$$\underset{n \times n}{X} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 2 & 1 & & \vdots \\ \vdots & \vdots & 2 & & 0 \\ 1 & n-1 & n-2 & & 1 \end{bmatrix}$$

① Least squares → Overfitting
(fitted values = data)

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

$$X\hat{\beta} = y \rightarrow \text{overfitting}$$

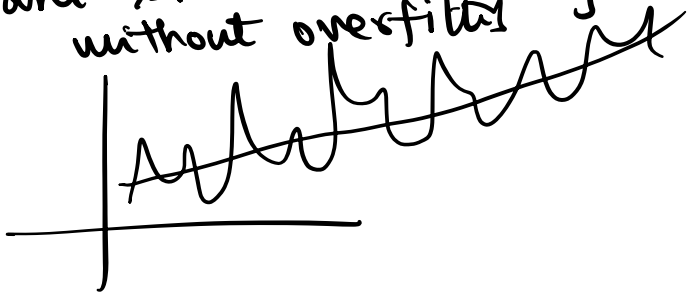
② Regularization:

Ridge: $\|y - X\beta\|^2 + \lambda \left[\sum_{j=2}^{n-1} \beta_j^2 \right]$

LASSO: $\|y - X\beta\|^2 + \lambda \sum_{j=2}^{n-1} |\beta_j|$

Why Regularize?

① Want smoother fits } Our preference
without overfitting



② Regularization leads to improved prediction.
(underlies CV for λ-selection)
Cross Validation

Formula for the Ridge Estimator

$$\|y - X\beta\|^2 + \lambda \sum_{j=2}^{n-1} \beta_j^2$$

$$(y - X\beta)^T (y - X\beta)$$

$$\nabla_\beta \left[\beta^T X^T X \beta - 2\beta^T X^T y + y^T y + \lambda \sum_{j=2}^{n-1} \beta_j^2 \right]$$

$$= 2 X^T X \beta - 2 X^T y + 2\lambda \begin{pmatrix} 0 \\ 0 \\ \beta_2 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = 0$$

$$\begin{pmatrix} 0 \\ 0 \\ \beta_2 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = J\beta \quad \text{where}$$

$$J = \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & \ddots \\ & & & & 1 \end{bmatrix}$$

$$X^T X \beta - X^T y + \lambda J \beta = 0$$

$$(X^T X + \lambda J)\beta = X^T y$$

$$\hat{\beta}^{\text{Ridge}}(\lambda) = (X^T X + \lambda I)^{-1} X^T y$$

$\lambda = 0$

$$\hat{\beta}^{\text{least square}} = (X^T X)^{-1} X^T y$$

Bayesian Regularization

$$y = X\beta + \varepsilon, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

likelihood: $\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{\|y - X\beta\|^2}{2\sigma^2}\right]$

$\beta\ (\beta_0, \beta_1, \ldots, \beta_{n-1})$ & σ

① $\beta_0, \beta_1, \ldots, \beta_{n-1} \overset{iid}{\sim} \text{Unif}(-C, C)$, $C \to \infty$

⇓ when $C \to \infty$

posterior

$$\beta \mid \underset{\sigma}{\text{data}} \sim N\left((X^T X)^{-1} X^T y,\ \sigma^2 (X^T X)^{-1}\right)$$

Homework 1

N: n-variate normal

② $\beta_0, \beta_1, \ldots, \beta_{n-1} \overset{iid}{\sim} N(0, C)$ (C large)

⇓ posterior

$\beta \mid \underset{\sigma}{\text{data}}$

$\frac{1}{\sqrt{2\pi}\,C} \exp\left[-\frac{\beta_j^2}{2C}\right]$ behaves like a constant

$$\sim N\left(\left(\frac{X^TX}{\sigma^2}+\frac{I}{C}\right)^{-1}\frac{X^Ty}{\sigma^2},\ \left(\frac{X^TX}{\sigma^2}+\frac{I}{C}\right)^{-1}\right)$$

Valid for every C (not just for large C)

for most β_j very similar to Unif $(-C, C)$

$$\frac{I\{-C<\beta_j<C\}}{2C}$$

If $C\to\infty$:

$$N\left(\left(\frac{X^TX}{\sigma^2}\right)^{-1}\frac{X^Ty}{\sigma^2},\ \left(\frac{X^TX}{\sigma^2}\right)^{-1}\right)$$

$$= N\left((X^TX)^{-1}X^Ty,\ \sigma^2(X^TX)^{-1}\right)$$

Coincides with posterior for Unif $(-\infty, \infty)$ prior.

$$\frac{1}{\sqrt{2\pi}\,C}\exp\left(-\frac{\beta_j^2}{2C}\right)$$

$\beta_j \in (-10, 10)$

$C = 10^{10}$

Summarize

① If you are using C large (either in Unif $(-C, C)$ or $N(0, C)$) the posterior mean = least squares will lead to overfitting.

---

③ $\beta_0, \beta_1 \overset{iid}{\sim} N(0, C)$, $\beta_2, \ldots, \beta_{n-1} \overset{iid}{\sim} N(0, \tau^2)$

C: large; for some small $\tau > 0$

This will lead to smooth fits without overfitting (depending on the chosen value of τ)

The formula for the posterior:

$$\beta \sim N_n\left(\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}, \underbrace{\begin{pmatrix}c & & & 0\\ & c & & \\ & & \tau^2 & \\ 0 & & & \ddots & \\ & & & & \tau^2\end{pmatrix}}_{Q}\right)$$

$$\beta \sim N(0, Q)$$

Check: posterior now becomes

$$N\left(\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}, \left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\right)$$

$\beta \mid \text{data}, \sigma$

posterior mean:

$$\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}$$

$$Q^{-1} = \begin{pmatrix}1/c & & & 0\\ & 1/c & & \\ & & 1/\tau^2 & \\ 0 & & & \ddots & \\ & & & & 1/\tau^2\end{pmatrix}$$

$$\approx \begin{pmatrix}0 & & & 0\\ & 0 & & \\ & & 1/\tau^2 & \\ 0 & & & \ddots & \\ & & & & 1/\tau^2\end{pmatrix} = \frac{1}{\tau^2}J$$

Posterior mean: $\left(\frac{X^TX}{\sigma^2} + \frac{1}{\tau^2}J\right)^{-1}\frac{X^Ty}{\sigma^2} = \left(X^TX + \frac{\sigma^2}{\tau^2}J\right)^{-1}X^Ty$

Ridge: $\left(X^TX + \lambda J\right)^{-1}X^Ty$

Coincide if $\boxed{\lambda = \frac{\sigma^2}{\tau^2}}$

$\beta_j \overset{iid}{\sim} N(0, C)$ (or Unif$(-C, C)$) C large

Bayesian = Frequentist

$(X^TX)^{-1}X^Ty$

$\beta_0, \beta_1 \overset{iid}{\sim} N(0, C)$, C large, $\beta_j \overset{iid}{\sim} N(0, \tau^2)$ $j = 2, \cdots, n-1$

Frequentist with Ridge Regularization = Bayesian

Prior 1 → $\beta_j \overset{iid}{\sim} N(0, C)$ → least square

Prior 2 → $\beta_0, \beta_1 \overset{iid}{\sim} N(0, C)$, $\beta_j \overset{iid}{\sim} N(0, \tau^2)$ → Ridge Regularizers

if τ is small

Prior 2 is restrictive.

Prior 3: $\beta_0, \beta_1 \overset{iid}{\sim} N(0, C)$, $\beta_j \overset{iid}{\sim} N(0, \tau^2)$

$\log \tau \sim \text{Unif}(-C, C)$, $\log \sigma \sim \text{Unif}(-C, C)$

More General

prior:

$$f(\beta, \sigma, \tau) \propto \frac{1}{\sqrt{\det Q}} \exp\left(-\frac{\beta^T Q^{-1} \beta}{2}\right) \frac{I\{-C < \log \tau < C\}}{\tau} \frac{I\{-C < \log \sigma < C\}}{\sigma}$$

$$\beta \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} C & & & \\ & C & & \\ & & \tau^2 & \\ & & & \ddots \end{pmatrix}\right) = N(0, Q) = \frac{\tau^{-1}\sigma^{-1}}{\sqrt{\det Q}} \exp\left(-\frac{\beta^T Q^{-1} \beta}{2}\right)$$

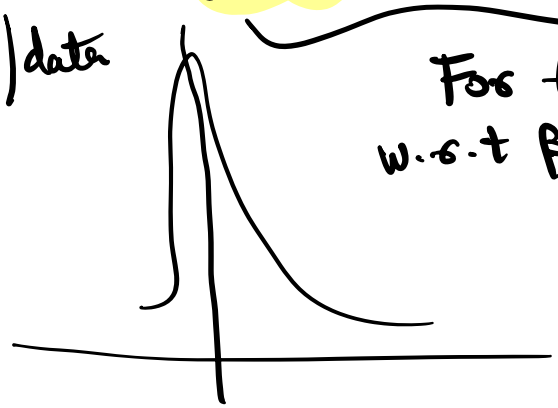
Likelihood: $\sigma^{-n} \exp\left(-\frac{\|y - X\beta\|^2}{2\sigma^2}\right)$

Posterior (β, τ, σ):

$$\frac{\sigma^{-n-1}\tau^{-1}}{\sqrt{\det Q}} \exp\left[-\frac{1}{2}\left(\frac{\|y - X\beta\|^2}{\sigma^2} + \beta^T Q^{-1} \beta\right)\right]$$

$f(\beta, \tau, \sigma \mid \text{data})$

For fixed σ, τ, integration w.r.t β is tractable.



$$\frac{\|y - X\beta\|^2}{\sigma^2} + \beta^T Q^{-1} \beta$$

$x^2 - 4x + 6 = (x-2)^2 + 2$

$$= (\beta - b)^T \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)(\beta - b) + \frac{y^T y}{\sigma^2} - b^T \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right) b$$

$$b = \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)^{-1} \frac{X^T y}{\sigma^2}$$

$f(\sigma, \tau)$ $\propto$ integral of $f(\beta, \sigma, \tau)$ w.r.t β

$\sigma, \tau \mid \text{data}$ $\beta, \sigma, \tau \mid \text{data}$

$$\propto \frac{\sigma^{-n-1} \tau^{-1}}{\sqrt{\det Q}} \sqrt{\det\left[\left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)^{-1}\right]} \exp\left(-\frac{y^T y}{2\sigma^2}\right) \exp\left[\frac{y^T X \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)^{-1} \frac{X^T y}{\sigma^2}}{2\sigma^2}\right]$$

1. Take a grid of values of σ & τ
2. Compute $\underbrace{f_{\sigma,\tau \mid \text{data}}}_{\text{posterior}}$ over the grid
3. Sample values of σ & τ from mean/mode of σ & τ
4. Sample β from $\beta \mid \underset{\sigma,\tau}{\text{data}}$: normal posterior

Why no overfitting?

$f_{data|\beta,\sigma}$: Likelihood

maximization leads to $\hat{\beta}$ = unregularized least squares, $\hat{\sigma} = 0$

$$f_{data|\tau,\sigma} = \int f_{data|\beta,\sigma}(data)\, f_{\beta|\tau}(\beta)\, d\beta$$

$f_{data|\tau,\sigma}$

$f_{\tau,\sigma|data}$

$$f_{\tau,\sigma|data} \propto f_{data|\tau,\sigma}\, f_{\tau,\sigma}$$

$f_{data|\tau,\sigma}$ ↕ $f_{data|\beta,\sigma}$