Ridge e Lasso em Regressão

STAT 224 — Aula 18

Adaptado do material de Yibi Huang

Compromisso Viés–Variância

Para um estimador $\hat{\beta}_j$,

$$\mathrm{MSE}(\hat{\beta}_j) = \mathbb{E}\big[(\hat{\beta}_j - \beta_j)^2\big] = \mathbb{E}\big[(\hat{\beta}_j - \mathbb{E}[\hat{\beta}_j])^2\big] + \big(\mathbb{E}[\hat{\beta}_j] - \beta_j\big)^2 = \underbrace{\mathrm{Var}(\hat{\beta}_j)}_{\text{variância}} + \underbrace{\big(\text{Viés de } \hat{\beta}_j\big)^2}_{\text{viés}^2}.$$

- Estimativas MMQ (OLS) para os β_j são **não enviesadas**.
- Contudo, as variâncias das estimativas OLS podem ser grandes quando
  - o número de preditores é grande; ou
  - os preditores são **multicolineares**.
- Podemos reduzir a variância de $\hat{\beta}_j$, possivelmente ao custo de um viés maior?

Regularização

- Estimativas OLS $\hat{\beta}_j$ não têm limite superior $\Rightarrow$ susceptíveis a variância muito alta.
- Ao **reduzir** as estimativas OLS $\hat{\beta}_j$ para 0, reduzimos substancialmente a variância, com aumento de viés muitas vezes negligenciável, melhorando a **previsão** futura.
- Em Aprendizagem Automática, "shrinkage" = **regularização**.
- Dois métodos comuns:
  - **Ridge**
  - **Lasso** (Least Absolute Shrinkage and Selection Operator).

OLS vs. Ridge vs. Lasso

OLS minimiza:

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}\right)^2.$$

Ridge minimiza a mesma soma com a restrição

$$\sum_{j=1}^{p}\hat{\beta}_j^2 \leq t.$$

Lasso minimiza a mesma soma com a restrição

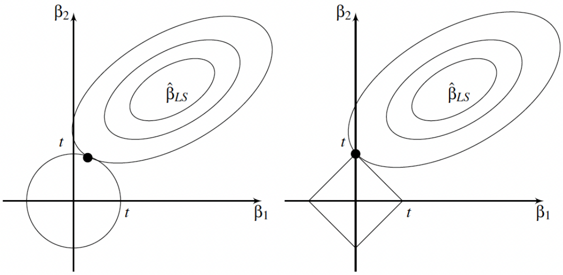
$$\sum_{j=1}^{p}\left|\hat{\beta}_j\right| \leq t.$$

Nota: não se impõe restrição sobre a magnitude do intercept $\hat{\beta}_0$.

Ilustração Geométrica de Ridge e Lasso

- As elipses são contornos de $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$, centradas em $(\hat{\beta}_{1,\mathrm{OLS}}, \hat{\beta}_{2,\mathrm{OLS}})$.
- **Esquerda:** a elipse intersecta o círculo de raio t na estimativa Ridge.
- **Direita:** a elipse intersecta o quadrado $(|\beta_1| + |\beta_2| \leq t)$ na estimativa Lasso.

Dica: escolher λ (ou t) pequeno o suficiente para obter estimativas estáveis.



Formas Equivalentes (Multiplicador de Lagrange)

Minimizar a soma de quadrados sob as restrições $\sum_{j=1}^{p} \hat{\beta}_j^2 \leq t$ ou $\sum_{j=1}^{p} |\hat{\beta}_j| \leq t$ é equivalente a:

$$\textbf{Ridge:} \quad \min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \textstyle\sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \textstyle\sum_{j=1}^{p} \beta_j^2,$$

$$\textbf{Lasso:} \quad \min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \textstyle\sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \textstyle\sum_{j=1}^{p} |\beta_j|.$$

Parâmetro de Contração λ (ou t)

- As estimativas $\hat{\beta}^{\text{Ridge}}_{j,\lambda}$ e $\hat{\beta}^{\text{Lasso}}_{j,\lambda}$ dependem de λ (ou t).
- λ controla a **contração** (tamanho) dos coeficientes.
- À medida que $\lambda \downarrow 0$ (ou $t \uparrow \infty$), Ridge/Lasso $\rightarrow$ OLS.
- À medida que $\lambda \uparrow \infty$ (ou $t \downarrow 0$), coeficientes $\rightarrow 0$ (modelo com intercept apenas).

Ridge e Lasso NÃO são invariantes à escala

Se mudarmos a unidade de um preditor X_j de polegadas para pés:

$$X_j' = X_j/12, \qquad \beta_j' = 12\,\beta_j,$$

então o produto $\beta_j' X_j' = \beta_j X_j$ permanece inalterado. Contudo, as estimativas Ridge/Lasso $\hat{\beta}'_{j,\lambda}$ não escalam de forma compatível (como $12\,\hat{\beta}_{j,\lambda}$), pois penalizam β_j grandes.

Padronizar preditores antes de aplicar Ridge/Lasso

Por convenção, padronizamos todos os preditores:

$$Z_j = \frac{X_j - \bar{X}_j}{s_j}, \qquad j = 1, \ldots, p,$$

onde s_j é o desvio-padrão amostral de X_j. Assume-se então que os X_j têm média 0 e variância 1 nas regressões Ridge/Lasso.

Ridge: estimativa e propriedades

- OLS: $\hat{\beta}_{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.
- Ridge: $\hat{\beta}_{\text{Ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.
- (Com $\boldsymbol{X}$ padronizada.) O valor esperado pode ser escrito como algo do tipo $(\boldsymbol{I}_p + \lambda \boldsymbol{X}^\top \boldsymbol{X})^{-1} \beta$ (indicando **viés**).

Caso simples ($\boldsymbol{X}^\top \boldsymbol{X} = I$, i.e., preditores padronizados e não correlacionados):

$$\hat{\beta}_{j,\text{Ridge}} = \frac{1}{1+\lambda} \hat{\beta}_{j,\text{OLS}}.$$

Variância de $\hat{\beta}_{j,\text{Ridge}}$ é tipicamente **menor** do que a de $\hat{\beta}_{j,\text{OLS}}$, sobretudo com multicolinearidade.

Propriedades do Lasso

- Não existe forma fechada para as estimativas Lasso.
- Também enviesado (para 0); variância geralmente menor que OLS.
- Pode não ser tão bom como Ridge na presença de forte multicolinearidade.
- Grande vantagem: **esparsidade**.

Esparsidade das Estimativas Lasso

Em modelos com muitos preditores,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

podemos acreditar que muitos $\beta_j = 0$. O Lasso devolve soluções esparsas: para λ suficientemente grande (ou t pequeno), alguns coeficientes passarão a zero. Logo, o Lasso realiza **seleção de variáveis**.

Como escolher λ?

- É necessário um critério *disciplinado* para escolher λ.
- Pretende-se minimizar o **erro quadrático médio**.
- Este tema liga-se ao problema mais amplo de **seleção de variáveis**.

Escolher λ por Validação Cruzada

- Dividir os dados em **treino** e **teste** (ou usar k-fold CV).
- Para cada λ: ajustar no treino, prever no teste e calcular o RMSE

$$\sqrt{\frac{1}{n_{\text{teste}}}\sum_{\text{teste}}(y_i - \hat{y}_i)^2}.$$

- Escolher o λ que minimiza o RMSE.
- As partições devem ser aleatórias; pode repetir várias vezes e usar a média do RMSE.