Aula 2: Regularização LASSO

Paulo C. Marques F.

Aula ministrada no Insper

12 de Fevereiro de 2016

Déjà vu: Regressão Linear Múltipla

- Para $i = 1, \ldots, n$, temos:
- Vetor de $p \geq 1$ preditoras $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$;
- Respostas $Y_i \in \mathbb{R}$;
- Variáveis aleatórias ϵ_i independentes e identicamente distribuídas (IID) com distribuição $\mathrm{N}(0, \sigma^2)$ (caso homoscedástico);
- Coeficientes de regressão $\beta = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$, em que β_0 é denominado "intercept".
- Os parâmetros são $(\beta, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_+$ (espaço paramétrico).
- Modelo de regressão linear múltipla:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i.$$

Verossimilhança

- Para $i = 1, \ldots, n$, observando que

$$Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} = \epsilon_i$$

são IID com distribuição $\mathrm{N}(0, \sigma^2)$, temos que a função de verossimilhança do modelo é dada por

$$\begin{aligned} L_y(\beta, \sigma^2) &= \prod_{i=1}^{n} f(y_i \mid \beta, \sigma^2) \\ &= \prod_{i=1}^{n} \left((2\pi)^{-1/2} \sigma^{-1} \exp\left(- \frac{(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij})^2}{2\sigma^2} \right) \right) \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp\left(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2 \right). \end{aligned}$$

Uma notação mais compacta e geométrica

- É conveniente trabalhar com a notação matricial:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{bmatrix}.$$

- A matriz $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ é denominada matriz de *design*.
- O modelo em notação matricial é $Y = \mathbf{X}\beta + \epsilon$.
- Note que $\mathrm{E}[Y] = \mathbf{X}\beta$.
- Com esta notação, reescrevemos a verossimilhança como

$$\begin{aligned} L_y(\beta, \sigma^2) &= (2\pi)^{-n/2}\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)\right) \\ &= (2\pi)^{-n/2}\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\|y - \mathbf{X}\beta\|^2\right), \end{aligned}$$

em que $\|u\| = \sqrt{u^T u}$ denota a norma euclidiana.

Máxima Verossimilhança e Mínimos Quadrados

- Para um σ^2 fixado, uma vez que a função logaritmo é crescente, encontrar $\hat{\beta}$ que maximize a verossimilhança $L_y(\beta, \sigma^2)$ é equivalente a maximizar $\log L_y(\beta, \sigma^2)$, o que por sua vez equivale a determinar

$$\hat{\beta} = \arg\min_{\beta} \|y - \mathbf{X}\beta\|^2.$$

- Portanto, para o modelo de regressão linear múltipla, a estimativa de máxima verossimilhança é exatamente a estimativa de mínimos quadrados ordinários ("ordinary least squares" ou OLS).
- Gauss já sabia disso tudo – e muitíssimo mais! – no final do século XVIII.

Solução geométrica

- Lembrando que rank $\mathbf{X}$ é o número de colunas linearmente independentes de $\mathbf{X}$, vamos assumir que rank $\mathbf{X} = p + 1$; o que implica que $\mathbf{X}^T\mathbf{X}$ é não singular.
- O ponto crucial é interpretar o produto $\mathbf{X}\beta$ como uma combinação linear das colunas de $\mathbf{X}$:

$$\mathbf{X}\beta = \beta_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} + \cdots + \beta_p \begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix}.$$

- Pela figura da lousa, vemos que para minimizar $\|y - \mathbf{X}\beta\|$ o vetor diferença $\delta = y - \mathbf{X}\beta$ deve ser perpendicular ao espaço coluna de $\mathbf{X}$, o que equivale a dizer que δ é ortogonal a cada uma das colunas de $\mathbf{X}$, ou seja, que $\mathbf{X}^T\delta = 0$.
- Portanto, $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T y$, e a solução de mínimos quadrados é

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

Detour: Bias-variance trade-off e Erro de Predição (1)

- Considere o caso mais geral: um oráculo escolhe uma função $\psi : \mathbb{R}^p \to \mathbb{R}$ e gera

$$Y_i = \psi(x_i) + \epsilon_i,$$

para $i = 1, \ldots, n$, em que os ϵ_i's são variáveis aleatórias IID com esperança 0 e variância σ^2.

- De posse dos dados $(x_1, y_1), \ldots, (x_n, y_n)$ gerados pelo oráculo, queremos construir uma função $\hat{\psi} : \mathbb{R}^d \to \mathbb{R}$ com a qual iremos predizer a resposta de um novo vetor de preditoras x_{n+1} por $\hat{\psi}(x_{n+1})$.
- A qualidade de $\hat{\psi}$ é avaliada na teoria clássica pelo seu erro quadrático médio ("mean squared error", ou MSE), definido por

$$\mathrm{MSE}[\hat{\psi}(x_{n+1})] = \mathrm{E}[(\psi(x_{n+1}) - \hat{\psi}(x_{n+1}))^2].$$

- Vale lembrar: em geral, não conhecemos ψ. Só conhecemos ψ se estivermos simulando os dados.

Detour: Bias-variance trade-off e Erro de Predição (2)

- Outra forma utilizada na teoria clássica para avaliar a qualidade de $\hat{\psi}$ é o seu erro de predição ("prediction error", ou PE), definido por

$$\text{PE}[\hat{\psi}(x_{n+1})] = \text{E}[(\, Y_{n+1} - \hat{\psi}(x_{n+1}))^2].$$

- Lembrando que $Y_{n+1} = \psi(x_{n+1}) + \epsilon_{n+1}$, a diferença entre MSE e PE é que na definição de PE levamos em conta o erro aleatório idiossincrático introduzido por ϵ_{n+1}.
- De fato, como avaliadores da qualidade de $\hat{\psi}(x_{n+1})$, o erro quadrático médio e o erro de predição são equivalentes, uma vez que podemos provar que

$$\text{PE}[\hat{\psi}(x_{n+1})] = \text{MSE}[\hat{\psi}(x_{n+1})] + \sigma^2.$$

Detour: Bias-variance trade-off e Erro de Predição (3)

- Com relação ao MSE, o resultado fundamental conhecido como "Bias-Variance trade-off" é que

$$\text{MSE}[\hat{\psi}(x_{n+1})] = \text{Bias}^2[\hat{\psi}(x_{n+1})] + \text{Var}[\hat{\psi}(x_{n+1})];$$

  em que

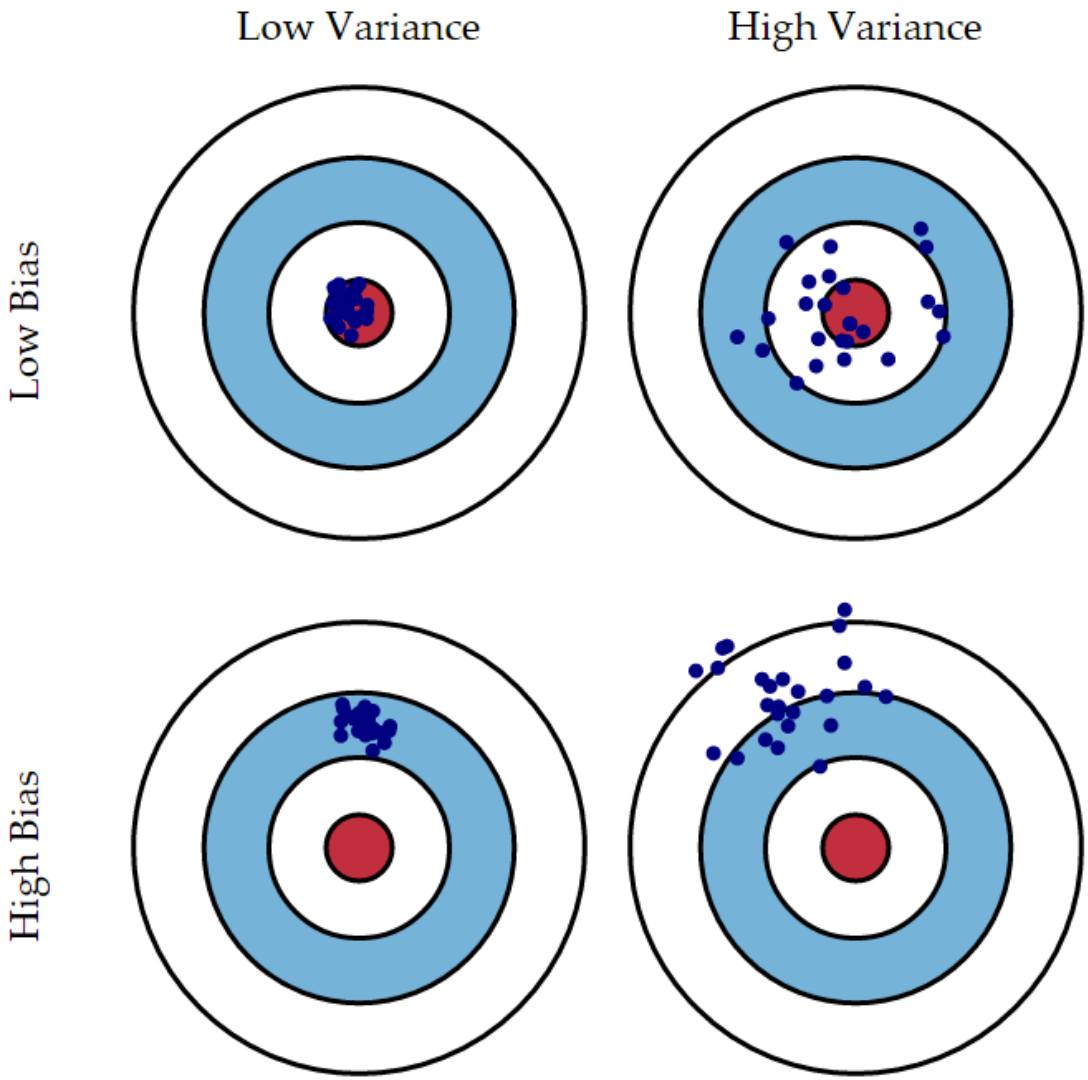
$$\text{Bias}[\hat{\psi}(x_{n+1})] = \text{E}[\hat{\psi}(x_{n+1})] - \psi(x_{n+1})$$

  e

$$\text{Var}[\hat{\psi}(x_{n+1})] = \text{E}[(\hat{\psi}(x_{n+1}) - \text{E}[\hat{\psi}(x_{n+1})])^2].$$

- Note que nesta decomposição se uma das parcelas crescer a outra necessariamente tem que diminuir.

Detour: Bias-variance trade-off e Erro de Predição (4)



Voltando para a regressão linear (1)

- Introduza a notação: $\mathbf{x}_{n+1} = \begin{bmatrix} 1 \\ x_{n+1,1} \\ \vdots \\ x_{n+1,p} \end{bmatrix}$.
- Para o modelo de regressão linear múltipla, a predição a partir da estimativa de mínimos quadrados é $\hat{\psi}(x_{n+1}) = \mathbf{x}_{n+1}^T \hat{\beta}$.
- Esta predição é não enviesada: $\mathrm{E}[\hat{\psi}(x_{n+1})] = \mathbf{x}_{n+1}^T \beta$.
- Muito importante: esta forma de predição clássica ignora a incerteza da estimativa pontual $\hat{\beta}$. Este é provavelmente um dos pontos de maior contraste com as soluções bayesianas que discutiremos nos seminários.
- Teorema de Gauss-Markov: a predição a partir da estimativa de mínimos quadrados tem a menor variância entre todas as previsões lineares não enviesadas.

Voltando para a regressão linear (2)

- Dois problemas com a predição no modelo de regressão linear a partir da estimativa de mínimos quadrados:
  - Não prevê bem – no sentido de que o erro de predição pode ser grande – principalmente quando temos muitas preditoras;
  - Não consegue reduzir ou zerar o valor de um subconjunto dos $\hat{\beta}_i$'s, o que pode dificultar a interpretação do modelo.
- Ao regularizar o modelo de regressão linear, vamos aceitar algum aumento no viés em troca de uma redução na variância.
- Também vamos buscar a compressão ("shrinkage") das estimativas com a finalidade de melhorar a interpretabilidade do modelo.

Regressão Ridge (1)

- As estimativas Ridge são as soluções

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left((y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \lambda \beta^T \beta \right),$$

  em que $\lambda \geq 0$ é denominado parâmetro de regularização.
- Se $\lambda = 0$, então $\hat{\beta}^{\text{ridge}}$ é igual à estimativa de mínimos quadrados.
- Quando $\lambda \to \infty$, todos os $\hat{\beta}_i$'s serão iguais a zero.
- Resultado:

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbb{I}_{p+1})^{-1} \mathbf{X}^T y.$$

Regressão Ridge (2)

Demonstração

Usando a notação em blocos para vetores em matrizes, defina

$$\tilde{y} = \begin{bmatrix} y \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{n+p+1} \quad \text{e} \quad \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbb{I}_{p+1} \end{bmatrix} \in \mathbb{R}^{(n+p+1)\times(p+1)}.$$

O resultado segue se observarmos que

$$\tilde{\mathbf{X}}^T\tilde{y} = \mathbf{X}^T y; \quad \tilde{\mathbf{X}}\beta = \begin{bmatrix} \mathbf{X}\beta \\ \sqrt{\lambda}\beta \end{bmatrix}; \quad \tilde{y} - \tilde{\mathbf{X}}\beta = \begin{bmatrix} y - \mathbf{X}\beta \\ \sqrt{\lambda}\beta \end{bmatrix};$$

$$(\tilde{y} - \tilde{\mathbf{X}}\beta)^T(\tilde{y} - \tilde{\mathbf{X}}\beta) = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) + \lambda\beta^T\beta;$$

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}_{p+1}.$$

Regressão Ridge (3)

- Suponha que temos uma situação de "colinearidade" em que duas das preditoras são perfeitamente correlacionadas.
- Isto implica que rank $\mathbf{X} < p + 1$.
- Ainda assim, teremos que rank $(\mathbf{X} + \lambda \mathbb{I}_{p+1}) = p + 1$.
- Esta é a motivação original de Hoerl e Kennard (1970).
- Detalhe importante: os praticantes centralizam e padronizam as preditoras.
- Estudos de simulação mostram que a regressão ridge melhora as predições, quando comparadas com as estimativas de mínimos quadrados, principalmente quando um grande número dos efeitos são pequenos.
- No entanto, a regressão ridge nunca zera os valores de um subconjunto dos $\hat{\beta}_i^{\text{ridge}}$'s. Não há vantagens interpretativas.
- Como escolher o valor de λ?

Regularização LASSO (1)

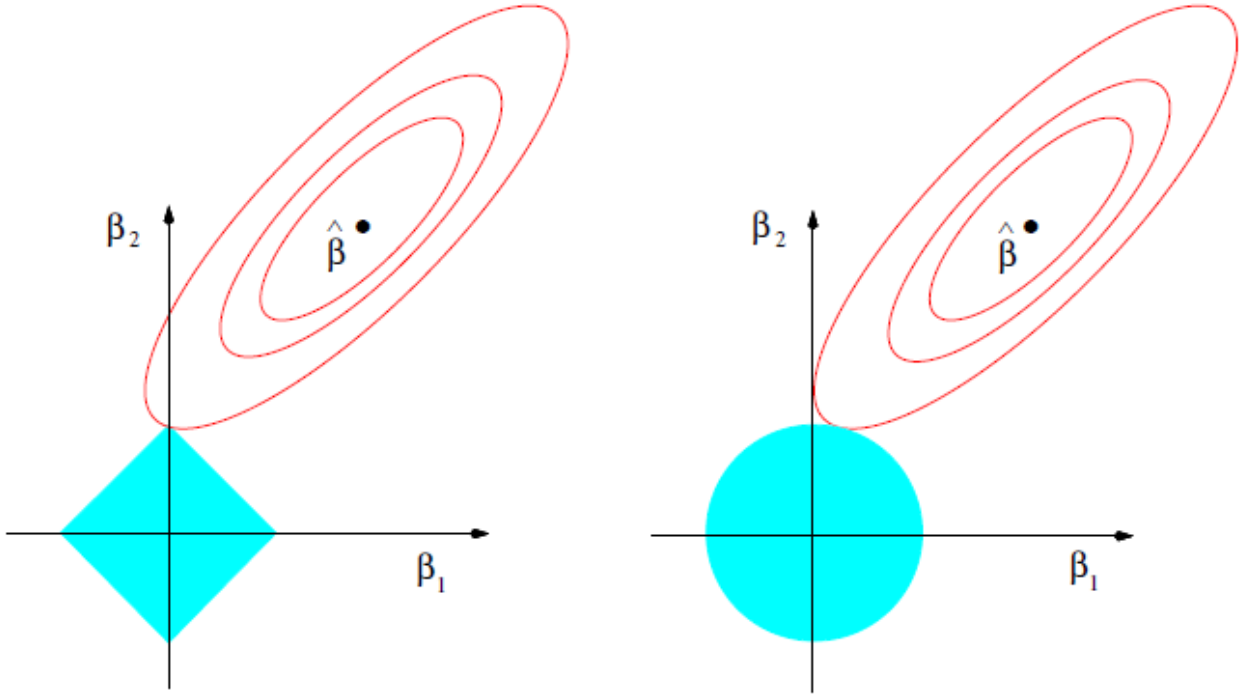
- As estimativas LASSO ("Least Absolute Selection and Shrinkage Operator") são as soluções

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left((y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \lambda \|\beta\|_1 \right),$$

  em que $\lambda \geq 0$ é o parâmetro de regularização e $\|u\|_1 = \sum_{j=1}^{p} |u_i|$ é a norma L_1.
- Novamente, os praticantes centralizam e padronizam as preditoras.
- As interpretações dos casos $\lambda = 0$ e $\lambda \to \infty$ são iguais às da regressão ridge.
- Simulações mostram que o LASSO melhora as predições, quando comparadas com as estimativas de mínimos quadrados, principalmente quando um número não muito grande dos efeitos são pequenos.

Regularização LASSO (2)

- A novidade é que o LASSO consegue zerar os valores de um subconjunto dos $\hat{\beta}_i^{\text{lasso}}$'s, melhorando a interpretação do modelo.



- Tanto na regressão ridge quanto no LASSO, o valor de λ é determinado minimizando a estimativa do erro de predição obtida por algum procedimento de validação cruzada.