Regularização e Regressão Ridge

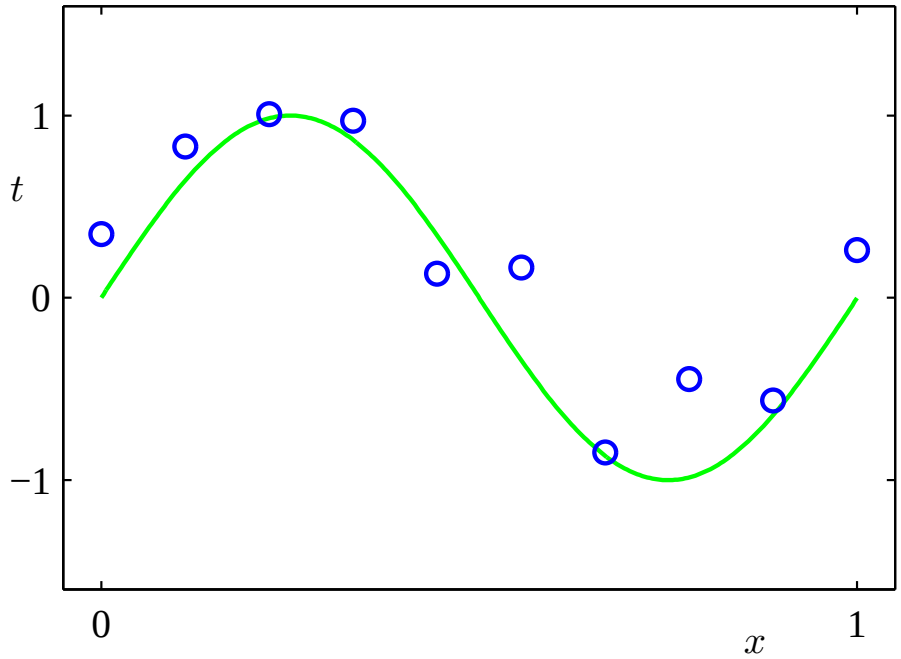
Prof. Dr. Leandro Balby Marinho

Análise de Dados II

Roteiro

Função Geradora

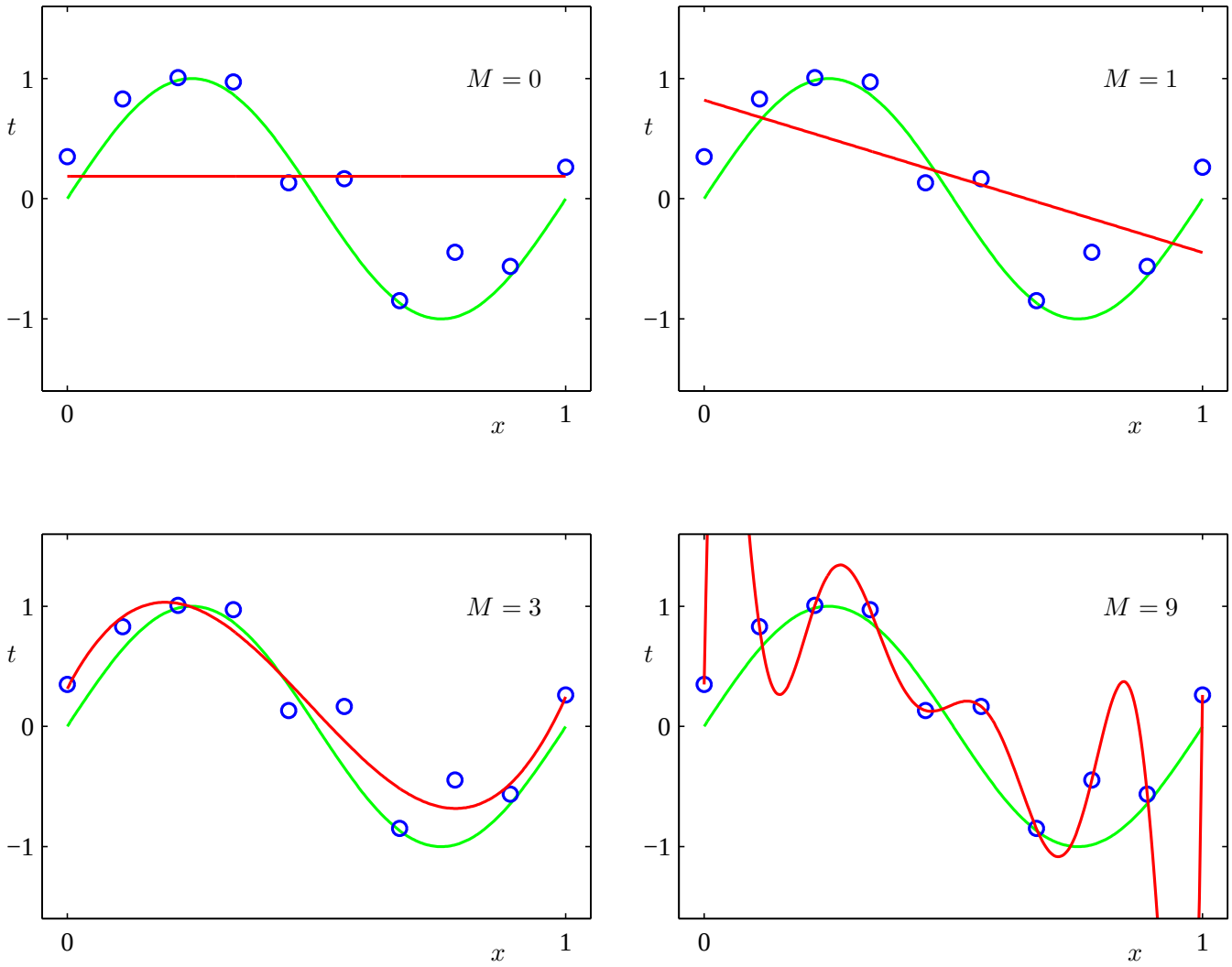
Considere os dados abaixo gerados pela função sen$(2\pi x)$ com ruído aleatório adicionado.



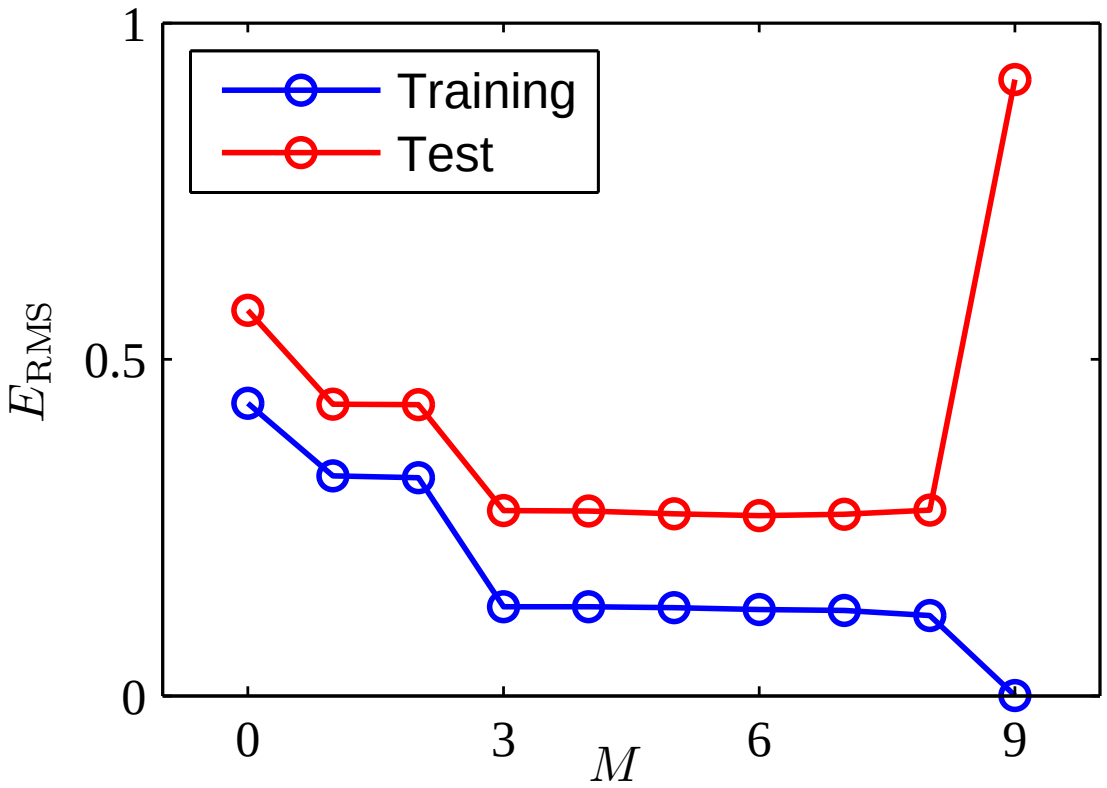
Normalmente, não sabemos o formato da função geradora, e então tentamos achar uma aproximação coerente.

Escolha do Modelo

Para $M =$ grau do polinômio:



Erro no Treino vs. Erro no Teste



Tamanho dos Parâmetros

Na regressão linear, overfitting é caracterizado por grandes valores dos parâmetros:

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0	0.19	0.82	0.31	0.35
w_1		−1.27	7.99	232.37
w_2			−25.43	−5321.83
w_3			17.37	48568.31
w_4				−231639.30
w_5				640042.26
w_6				−1061800.52
w_7				1042400.18
w_8				−557682.99
w_9				125201.43

Formato Desejado do Custo

Queremos balancear:

i. Ajuste aos dados
ii. Magnitude dos coeficientes (Complexidade do modelo)

Custo total=medida do ajuste (RSS) + medida da magnitude.

Medida da magnitude dos coeficientes de regressão

Que medida é indicativa da magnitude dos coeficientes:

- Soma:

$$\sum_{j=0}^{D} w_j$$

- Soma de valores absolutos (norma L_1):

$$\sum_{j=0}^{D} |w_j| = ||\mathbf{w}||_1$$

- Soma dos quadrados (normal L_2):

$$\sum_{j=0}^{D} {w_j}^2 = ||\mathbf{w}||_2^2$$

Regressão Ridge (também chamada de Regularização L2)

$$\textbf{Custo total} = \underbrace{\text{medida do ajuste}}_{\text{RSS}(\mathbf{w})} + \underbrace{\text{medida da magnitude}}_{||\mathbf{w}||_2^2}$$

Tarefa: Selecionar $\hat{\mathbf{w}}$ para minimizar

$$\text{RSS}(\hat{\mathbf{w}}) + \lambda ||\hat{\mathbf{w}}||_2^2$$

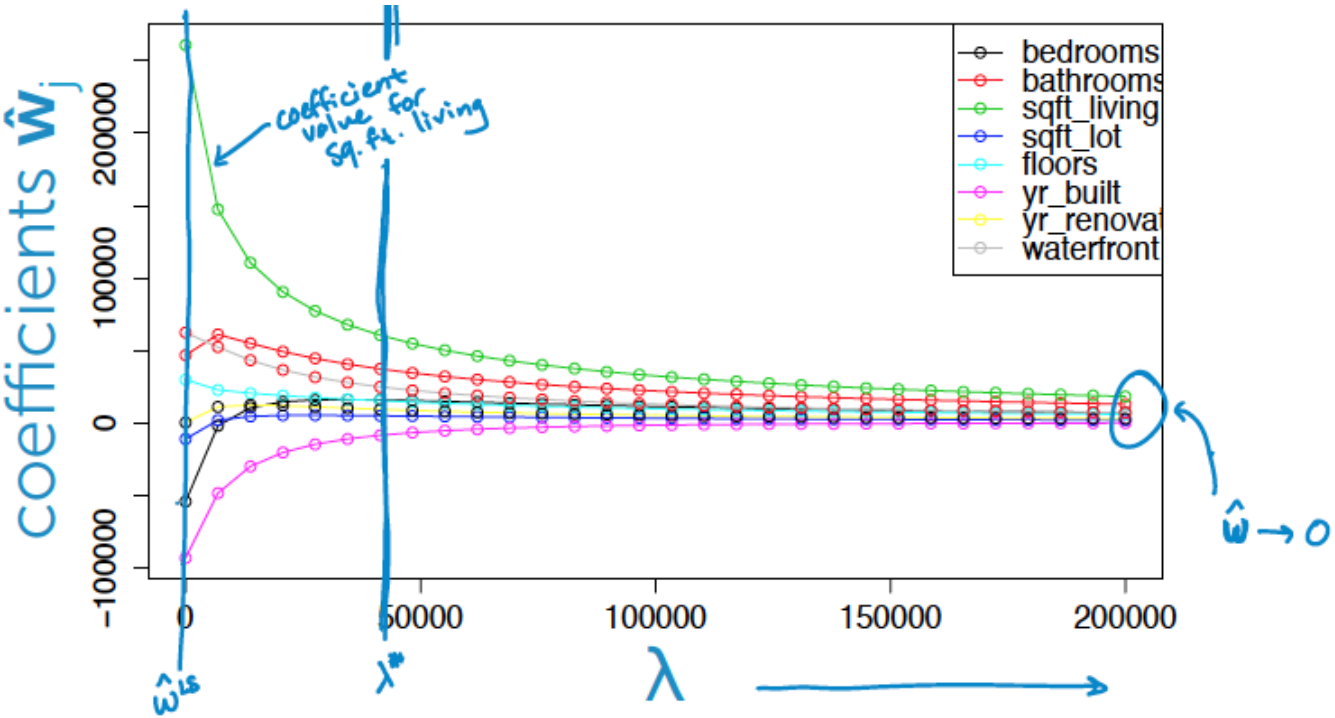
- $\lambda = 0$: problema reduz a achar os mínimos quadrados ($||\hat{\mathbf{w}}^{\text{MQ}}||_2^2$).
- $\lambda = \infty$
  - se $\hat{\mathbf{w}} \neq 0$ o custo total é ∞
  - se $\hat{\mathbf{w}} = 0$ o custo total é RSS(0)
- $0 < \lambda < \infty$:

$$0 \leq ||\hat{\mathbf{w}}||_2^2 \leq ||\hat{\mathbf{w}}^{\text{MQ}}||_2^2$$

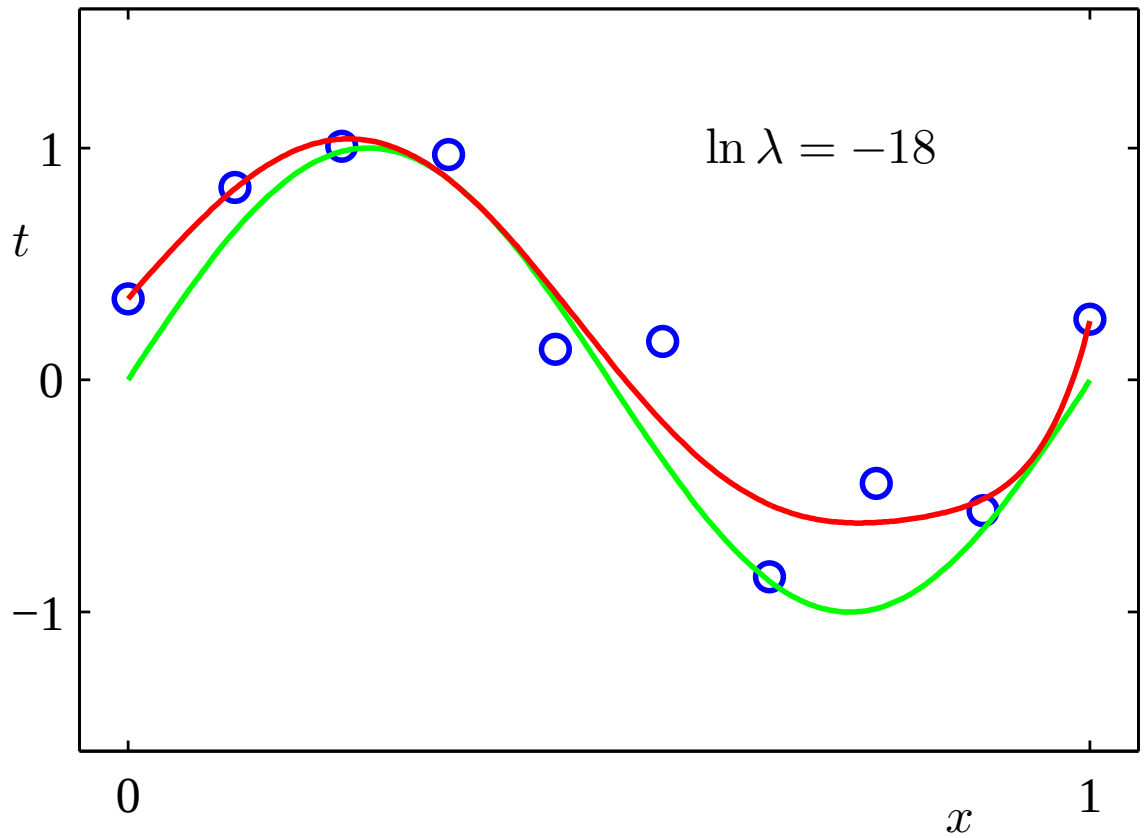
Trade-off Bias-Variância

- λ grande:
  - Bias grande, baixa variância
  - Exemplo: $\hat{\mathbf{w}} = 0$ para $\lambda = \infty$
- λ pequeno:
  - Bias pequeno, alta variância
  - Exemplo: Método dos mínimos quadrados para um polinômio de alta ordem para $\lambda = 0$

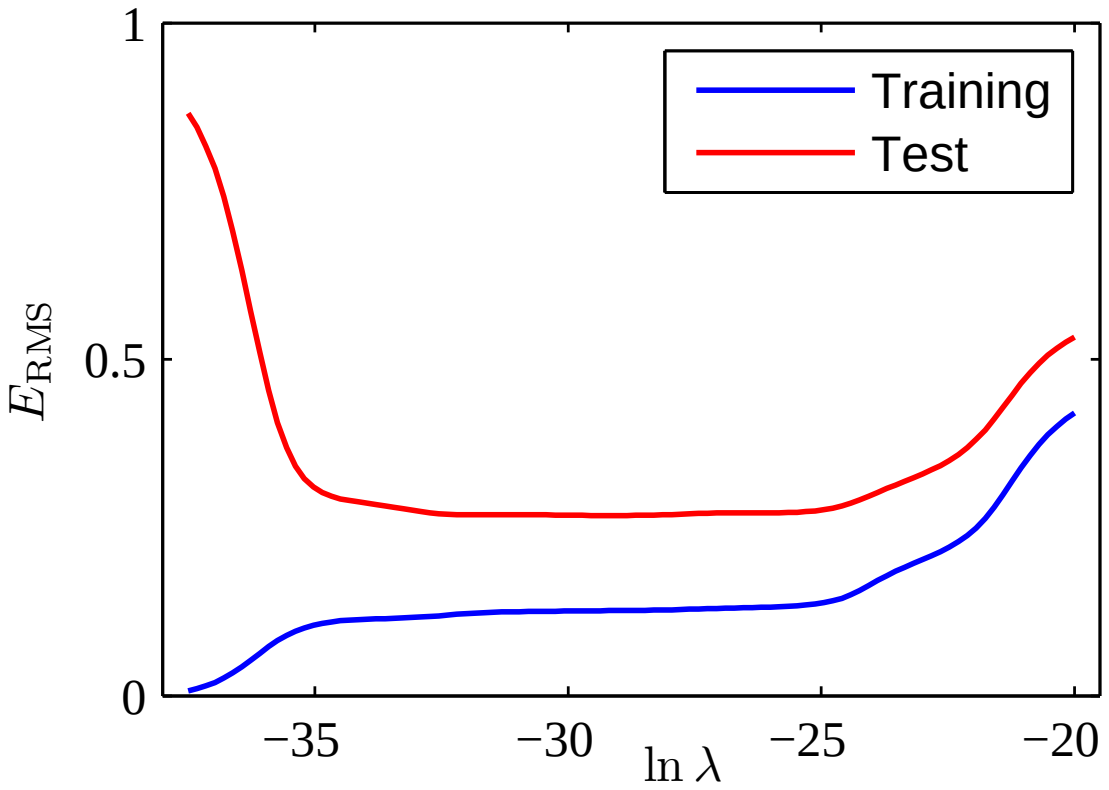
Caminho dos Coeficientes na Regressão Ridge



Impacto da Regularização

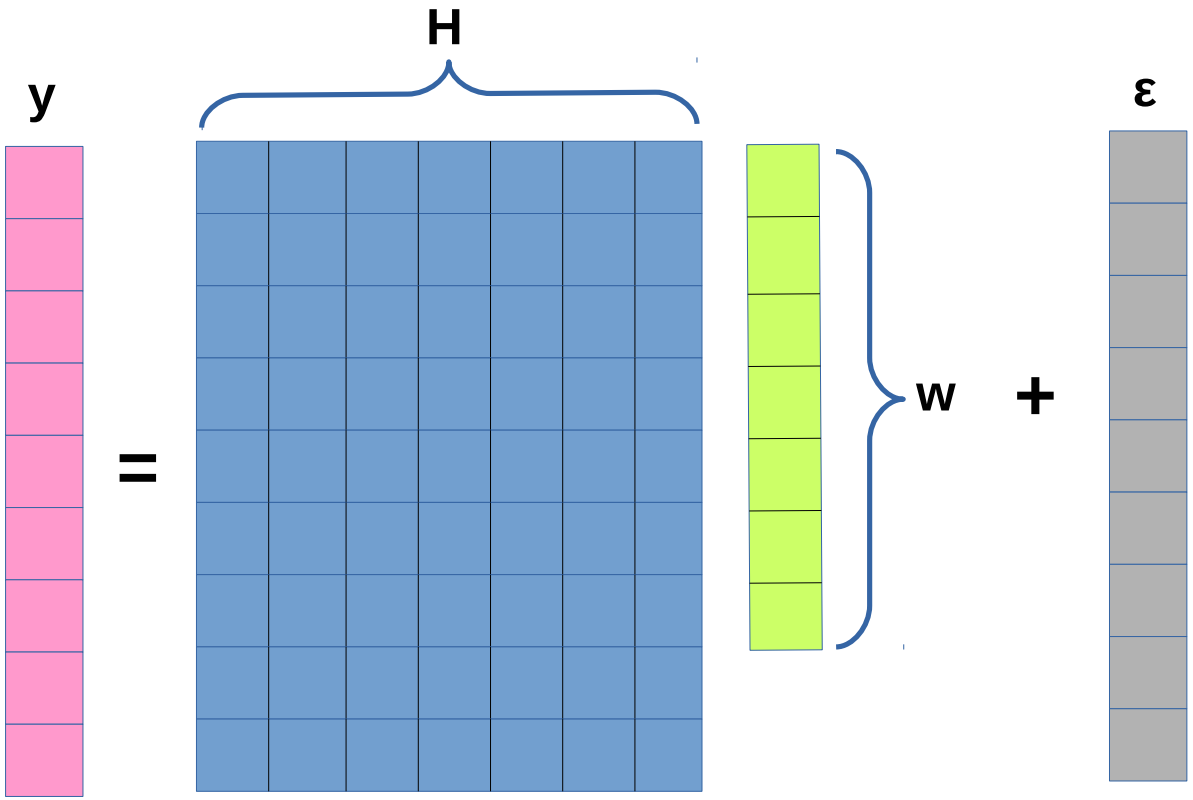


Impacto da Regularização no Teste



Roteiro

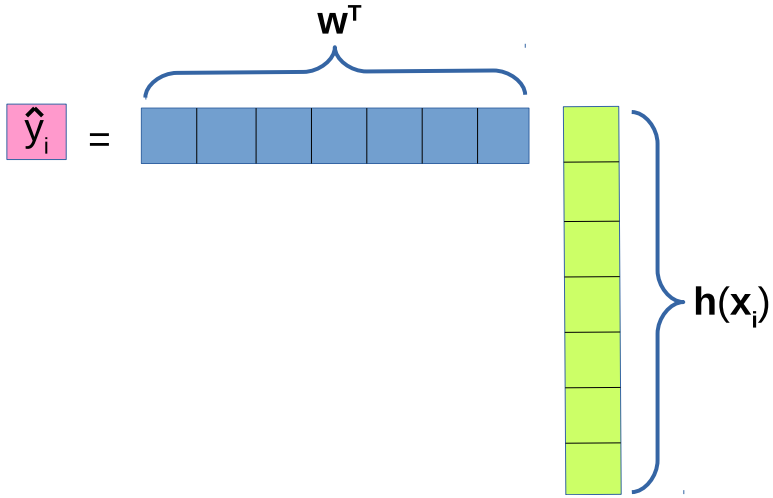
Usando notação de matrizes: todas as observações



$$\mathbf{y} = \mathbf{H}\mathbf{w} + \epsilon$$

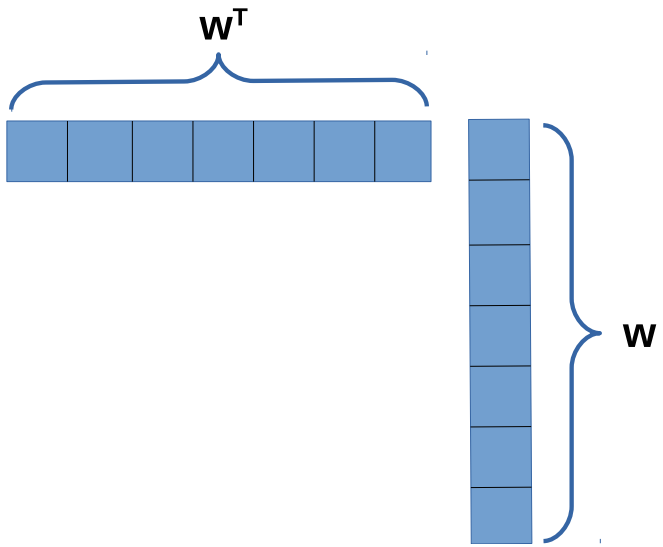
Custo de uma curva D-dimensional

$$\begin{aligned} \text{RSS}(\mathbf{w}) &= \sum_{i=1}^{N} (y^{(i)} - \mathbf{w}^T \mathbf{h}(\mathbf{x}^{(i)}))^2 \\ &= (\mathbf{y} - \mathbf{H}\mathbf{w})^T (\mathbf{y} - \mathbf{H}\mathbf{w}) \end{aligned}$$



Notação vetorial para $||\mathbf{w}||_2^2$

$$||\mathbf{w}||_2^2 = w_0^2 + w_1^2 + w_2^2 + \ldots + w_D^2$$



$$||\mathbf{w}||_2^2 = \mathbf{w}^T\mathbf{w}$$

Custo da regressão Ridge

Na forma matricial o custo agora é dado por:

$$\text{RSS}(\mathbf{w}) + \lambda||\mathbf{w}||_2^2 = (\mathbf{y} - \mathbf{H}\mathbf{w})^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w}$$

Gradiente do RSS

$$\nabla \left[\text{RSS}(\mathbf{w}) + \lambda ||\mathbf{w}||_2^2 \right] = \nabla \left[(\mathbf{y} - \mathbf{H}\mathbf{w})^T (\mathbf{y} - \mathbf{H}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w} \right]$$

Gradiente do RSS

$$\begin{aligned}\nabla \left[\text{RSS}(\mathbf{w}) + \lambda ||\mathbf{w}||_2^2\right] &= \nabla \left[(\mathbf{y} - \mathbf{H}\mathbf{w})^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right] \\ &= \nabla \left[(\mathbf{y} - \mathbf{H}\mathbf{w})^T(\mathbf{y} - \mathbf{H}\mathbf{w})\right] + \nabla \left[\lambda \mathbf{w}^T\mathbf{w}\right]\end{aligned}$$

Gradiente do RSS

$$\begin{aligned}
\nabla \left[\text{RSS}(\mathbf{w}) + \lambda ||\mathbf{w}||_2^2\right] &= \nabla \left[(\mathbf{y} - \mathbf{H}\mathbf{w})^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right] \\
&= \nabla \left[(\mathbf{y} - \mathbf{H}\mathbf{w})^T(\mathbf{y} - \mathbf{H}\mathbf{w})\right] + \nabla \left[\lambda \mathbf{w}^T\mathbf{w}\right] \\
&= -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda\mathbf{w}
\end{aligned}$$

Gradiente do RSS

$$
\begin{aligned}
\nabla \left[\text{RSS}(\mathbf{w}) + \lambda ||\mathbf{w}||_2^2\right] &= \nabla \left[(\mathbf{y} - \mathbf{H}\mathbf{w})^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right] \\
&= \nabla \left[(\mathbf{y} - \mathbf{H}\mathbf{w})^T(\mathbf{y} - \mathbf{H}\mathbf{w})\right] + \nabla \left[\lambda \mathbf{w}^T\mathbf{w}\right] \\
&= -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda \mathbf{w} \\
&= -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda \mathbf{I}\mathbf{w} \quad (\mathbf{I} = \text{identidade})
\end{aligned}
$$

Calculando parâmetros de forma fechada

$$\nabla \text{RSS}(\mathbf{w}) = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda\mathbf{I}\mathbf{w} = 0$$

Resolvendo para $\mathbf{w}$:

$$-2\mathbf{H}^T\mathbf{y} + 2\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + 2\lambda\mathbf{I}\hat{\mathbf{w}} = 0$$

Calculando parâmetros de forma fechada

$$\nabla \text{RSS}(\mathbf{w}) = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda\mathbf{I}\mathbf{w} = 0$$

Resolvendo para $\mathbf{w}$:

$$-2\mathbf{H}^T\mathbf{y} + 2\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + 2\lambda\mathbf{I}\hat{\mathbf{w}} = 0$$

$$-\mathbf{H}^T\mathbf{y} + \mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + \lambda\mathbf{I}\hat{\mathbf{w}} = 0 \quad \text{(divide ambos os lados por 2)}$$

Calculando parâmetros de forma fechada

$$\nabla \text{RSS}(\mathbf{w}) = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda\mathbf{I}\mathbf{w} = 0$$

Resolvendo para $\mathbf{w}$:

$$-2\mathbf{H}^T\mathbf{y} + 2\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + 2\lambda\mathbf{I}\hat{\mathbf{w}} = 0$$

$$-\mathbf{H}^T\mathbf{y} + \mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + \lambda\mathbf{I}\hat{\mathbf{w}} = 0 \quad \text{(divide ambos os lados por 2)}$$

$$\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + \lambda\mathbf{I}\hat{\mathbf{w}} = \mathbf{H}^T\mathbf{y}$$

Calculando parâmetros de forma fechada

$$\nabla \text{RSS}(\mathbf{w}) = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda\mathbf{I}\mathbf{w} = 0$$

Resolvendo para $\mathbf{w}$:

$$-2\mathbf{H}^T\mathbf{y} + 2\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + 2\lambda\mathbf{I}\hat{\mathbf{w}} = 0$$

$$-\mathbf{H}^T\mathbf{y} + \mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + \lambda\mathbf{I}\hat{\mathbf{w}} = 0 \quad \text{(divide ambos os lados por 2)}$$

$$\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + \lambda\mathbf{I}\hat{\mathbf{w}} = \mathbf{H}^T\mathbf{y}$$

$$(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})\hat{\mathbf{w}} = \mathbf{H}^T\mathbf{y}$$

Calculando parâmetros de forma fechada

$$\nabla \text{RSS}(\mathbf{w}) = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) + 2\lambda\mathbf{I}\mathbf{w} = 0$$

Resolvendo para $\mathbf{w}$:

$$\begin{aligned}
-2\mathbf{H}^T\mathbf{y} + 2\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + 2\lambda\mathbf{I}\hat{\mathbf{w}} &= 0 \\
-\mathbf{H}^T\mathbf{y} + \mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + \lambda\mathbf{I}\hat{\mathbf{w}} &= 0 \quad \text{(divide ambos os lados por 2)} \\
\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} + \lambda\mathbf{I}\hat{\mathbf{w}} &= \mathbf{H}^T\mathbf{y} \\
(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})\hat{\mathbf{w}} &= \mathbf{H}^T\mathbf{y} \\
\hat{\mathbf{w}} &= (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}
\end{aligned}$$

Custo da inversão de matrizes (quando inversível): $O(D^3)$

Calculando parâmetros de forma fechada)

$$\hat{\mathbf{w}}^{\text{ridge}} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$$

- $\lambda = 0$: problema reduz a achar os mínimos quadrados ($(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$).
- $\lambda = \infty$: $\hat{\mathbf{w}}^{\text{ridge}} = 0$ (análogo a dividir por infinito)

Gradiente Descendente

Gradient-Descent-Ridge

1 **while** not converged

2 $$w_j^{(t+1)} = w_j^{(t)} - \alpha \left[-2 \sum_{i=1}^{N} h_j(\mathbf{x}^{(i)})(y^{(i)} - \hat{y}^{(i)}) + 2\lambda w_j^{(t)} \right]$$

Gradiente Descendente

Gradient-Descent-Ridge

1 **while** not converged

2 $$w_j^{(t+1)} = (1 - 2\alpha\lambda) w_j^{(t)} + 2\alpha \sum_{i=1}^{N} h_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \hat{y}^{(i)}\right)$$

Gradient-Descent-OLS

1 **while** not converged

2 $$w_j^{(t+1)} = w_j^{(t)} + 2\alpha \sum_{i=1}^{N} h_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \hat{y}^{(i)}\right)$$

OLS = Ordinary Least Squares (Mínimos Quadrados Ordinário)

$\hat{y}^{(i)} = \mathbf{h}(\mathbf{x}^{(i)})^T \mathbf{w}$

Roteiro

Como encontrar o melhor λ?

- Dividindo os dados em K blocos diferentes de forma aleatória.
- No caso de uma partição $(1/2, 1/2)$ dos dados,
  1. A primeira parte é usada para treino e a segunda para validação.
  2. A segunda parte é usada para treino e a primeira para validação.
- Essa ideia pode ser generalizada para k partições de igual tamanho.
- Em cada execução, $k - 1$ partições são usadas para treino e uma para validação.
- O procedimento é repetido k vezes e a média do MSE calculada.

Validação Cruzada 5-fold

For $k = 1, \ldots, K$

1. Estime $\hat{\mathbf{w}}_{\lambda}^{(k)}$ nos blocos de treino
2. Calcule o erro no bloco de validação $\mathcal{L}_k(\lambda)$

Calcule erro médio: $\text{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_k(\lambda)$ (repita para todos os λ)

Referências

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, 2013.

Emily Fox and Carlos Guestrin. Machine Learning Specialization. Curso online disponível em `https://www.coursera.org/specializations/machine-learning`. Último acesso: 14/09/2017.