

Seleção de Atributos

Prof. Dr. Leandro Balby Marinho



Aprendizagem de Máquina

Roteiro

1. Busca Exaustiva/Gulosa

2. Regressão Lasso

3. Solução para o LASSO

Por que fazer seleção de atributos?

Eficiência:

- ▶ Se $\text{size}(\hat{\mathbf{w}})=100\text{B}$, cada predição fica cara.
- ▶ Se $\hat{\mathbf{w}}$ é esparsos, então cálculos só dependem das entradas $\neq 0$.

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

Interpretabilidade: Quais atributos são mais relevantes para a predição?

Todos os subconjuntos

- ▶ Para cada subconjunto possível calcule o erro e escolha o de menor erro.
- ▶ Para escolher a complexidade do modelo:
 1. Use um conjunto de validação.
 2. Use validação cruzada.
 3. Outras métricas como BIC.

Complexidade: Para D atributos temos 2^{D+1} subconjuntos (lembre do w_0).

Abordagem Gulosa

Forward Selection:

1. Comece com o conjunto vazio de atributos $F_0 = \emptyset$
2. Calcule o ajuste do modelo usando o conjunto atual de atributos F_t para obter $\hat{\mathbf{w}}^{(t)}$
3. Selecione o melhor próximo atributo $h_{j^*}(\mathbf{x})$
4. $F_{t+1} \rightarrow F_t \cup \{h_{j^*}(\mathbf{x})\}$
5. Chame o procedimento recursivamente passando F_t
6. **Itere até que nenhum novo atributo traga ganho significativo em relação ao modelo atual.**

Complexidade: $O(D^2) \ll O(2^{D+1})$

Roteiro

1. Busca Exaustiva/Gulosa

2. Regressão Lasso

3. Solução para o LASSO

Usando Regularização para Seleção de Atributos

- ▶ Comece com o modelo completo (todos os atributos possíveis)
- ▶ Reduza alguns coeficientes a zero.
- ▶ Coeficientes diferentes de zero são os selecionados.

Regressão Lasso

$$\text{Custo total} = \underbrace{\text{medida do ajuste}}_{\text{RSS}(\mathbf{w})} + \underbrace{\text{medida da magnitude}}_{\|\mathbf{w}\|_1}$$

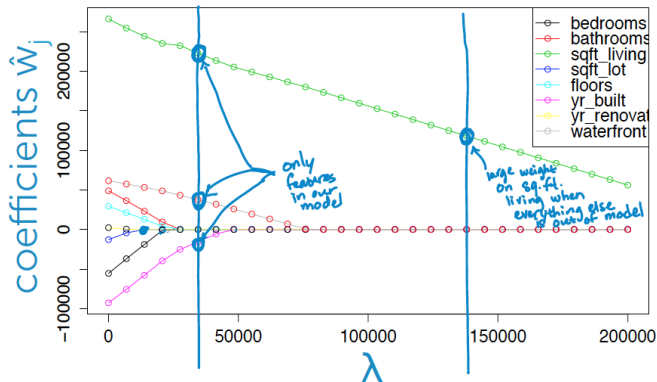
Tarefa: Selecionar $\hat{\mathbf{w}}$ para minimizar

$$\text{RSS}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$$

- ▶ $\lambda = 0$: problema reduz a achar os mínimos quadrados.
- ▶ $\lambda = \infty$: $\hat{\mathbf{w}} = 0$
- ▶ $0 < \lambda < \infty$:

$$0 \leq \|\hat{\mathbf{w}}\|_1 \leq \|\hat{\mathbf{w}}^{\text{MQ}}\|_1$$

Exemplo



Custo Regressão Lasso

$$\text{Custo total} = \text{RSS}(\mathbf{w}) + \lambda \sum_{j=0}^N |w_j|$$

Problemas:

- ▶ Quais as derivadas de $|w_j|$?
- ▶ Mesmo que pudéssemos calcular todas as derivadas, não há solução fechada.

Coordinate Descent

- ▶ **Objetivo:** $\min_{\mathbf{w}} g(\mathbf{w})$
- ▶ Geralmente difícil para todas as coordenadas, mas **simples para cada coordenada**.

Coordinate-Descent

- 1 initialize \mathbf{w}
- 2 **while** not converged
- 3 pick a coordinate j
- 4 $\hat{w}_j = \min_{\omega} g(w_0, \dots, w_{j-1}, \omega, w_{j+1}, \dots, w_D)$

Coordinate Descent

- ▶ Como escolhemos a próxima coordenada?
 - ▶ De forma aleatória, round robin, ...
- ▶ Não precisa escolher tamanho do passo da descida.
- ▶ Útil para muitos problemas:
 - ▶ Converge para o ótimo em alguns casos (e.g. funções fortemente convexas).
 - ▶ Converge para a função objetivo do Lasso.

Normalização de Features

- ▶ Aplique a normalização nas colunas (não nas linhas):

- ▶
$$\underline{h}_j(\mathbf{x}^{(k)}) = \frac{h_j(\mathbf{x}^{(k)})}{\sqrt{\sum_{i=1}^N h_j(\mathbf{x}^{(i)})^2}}$$

- ▶ Aplique o mesmo fator de normalização aos atributos de teste:

- ▶
$$\underline{h}_j(\mathbf{x}^{(k)}) = \frac{h_j(\mathbf{x}^{(k)})}{\sqrt{\sum_{i=1}^N h_j(\mathbf{x}^{(i)})^2}}$$



Mínimos Quadrados com Coordinate Descent

- ▶ $\text{RSS}(\mathbf{w}) = \sum_{i=1}^N \left(y^{(i)} - \sum_{j=0}^D w_j h_j(\mathbf{x}^{(i)}) \right)^2$
- ▶ Fixe todas as coordenadas \mathbf{w}_{-j} e calcule a derivada com relação à w_j

Mínimos Quadrados com Coordinate Descent

► $\text{RSS}(\mathbf{w}) = \sum_{i=1}^N \left(y^{(i)} - \sum_{j=0}^D w_j \underline{h}_j(\mathbf{x}^{(i)}) \right)^2$

► Fixe todas as coordenadas \mathbf{w}_{-j} e calcule a derivada com relação à w_j

$$\frac{\partial}{\partial w_j} \text{RSS}(\mathbf{w}) = -2 \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{j=0}^D w_j \underline{h}_j(\mathbf{x}^{(i)}) \right)$$

Mínimos Quadrados com Coordinate Descent

► $\text{RSS}(\mathbf{w}) = \sum_{i=1}^N \left(y^{(i)} - \sum_{j=0}^D w_j \underline{h}_j(\mathbf{x}^{(i)}) \right)^2$

► Fixe todas as coordenadas \mathbf{w}_{-j} e calcule a derivada com relação à w_j

$$\begin{aligned} \frac{\partial}{\partial w_j} \text{RSS}(\mathbf{w}) &= -2 \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{j=0}^D w_j \underline{h}_j(\mathbf{x}^{(i)}) \right) \\ &= -2 \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{k \neq j} w_k \underline{h}_k(\mathbf{x}^{(i)}) - w_j \underline{h}_j(\mathbf{x}^{(i)}) \right) \end{aligned}$$

Mínimos Quadrados com Coordinate Descent

- ▶ $\text{RSS}(\mathbf{w}) = \sum_{i=1}^N \left(y^{(i)} - \sum_{j=0}^D w_j \underline{h}_j(\mathbf{x}^{(i)}) \right)^2$
- ▶ Fixe todas as coordenadas \mathbf{w}_{-j} e calcule a derivada com relação à w_j

$$\begin{aligned} \frac{\partial}{\partial w_j} \text{RSS}(\mathbf{w}) &= -2 \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{j=0}^D w_j \underline{h}_j(\mathbf{x}^{(i)}) \right) \\ &= -2 \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{k \neq j} w_k \underline{h}_k(\mathbf{x}^{(i)}) - w_j \underline{h}_j(\mathbf{x}^{(i)}) \right) \\ &= -2 \underbrace{\sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{k \neq j} w_k \underline{h}_k(\mathbf{x}^{(i)}) \right)}_{\triangleq \rho_j} + 2 w_j \underbrace{\sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)})^2}_{=1} \end{aligned}$$

Mínimos Quadrados com Coordinate Descent

- ▶ $\text{RSS}(\mathbf{w}) = \sum_{i=1}^N \left(y^{(i)} - \sum_{j=0}^D w_j h_j(\mathbf{x}^{(i)}) \right)^2$
- ▶ Iguale derivada a 0 e resolva para w_j :

$$-2p_j + 2\hat{w}_j = 0$$

$$\hat{w}_j = p_j$$

Mínimos Quadrados com Coordinate Descent

Coordinate-Descent-OLS

- 1 initialize $\hat{\mathbf{w}}$
- 2 **while** not converged
- 3 pick a coordinate j
- 4
$$p_j = \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)})(y^{(i)} - \hat{y}^{(i)}(\hat{\mathbf{w}}_{-j}))$$
- 5 $\hat{w}_j = p_j$

Roteiro

1. Busca Exaustiva/Gulosa
2. Regressão Lasso
3. Solução para o LASSO

Otimização do Objetivo Lasso

$$\text{RSS}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 = \sum_{i=1}^N \left(y^{(i)} - \sum_{j=0}^D w_j h_j(\mathbf{x}^{(i)}) \right)^2 + \lambda \sum_{j=0}^D |w_j|$$

- Fixe todas as coordenadas w_{-j} e calcule a derivada parcial com relação à w_j .

Parte 1: Derivada parcial do RSS

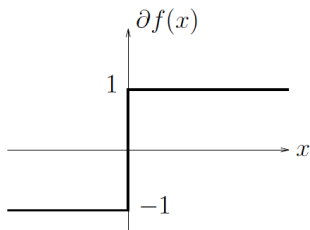
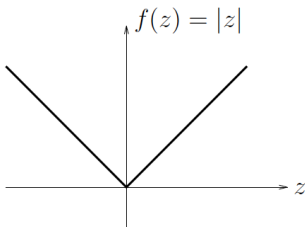
$$\blacktriangleright \text{RSS}(\mathbf{w}) = \sum_{i=1}^N \left(y^{(i)} - \sum_{j=0}^D w_j h_j(\mathbf{x}^{(i)}) \right)^2$$

$$\begin{aligned} \frac{\partial}{\partial w_j} \text{RSS}(\mathbf{w}) &= -2 \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{j=0}^D w_j \underline{h}_j(\mathbf{x}^{(i)}) \right) \\ &= -2 \sum_{i=1}^N \underline{h}_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{k \neq j} w_k \underline{h}_k(\mathbf{x}^{(i)}) - w_j \underline{h}_j(\mathbf{x}^{(i)}) \right) \\ &= -2 \underbrace{\sum_{i=1}^N h_j(\mathbf{x}^{(i)}) \left(y^{(i)} - \sum_{k \neq j} w_k \underline{h}_k(\mathbf{x}^{(i)}) \right)}_{\triangleq \rho_j} + 2w_j \underbrace{\sum_{i=1}^N h_j(\mathbf{x}^{(i)})^2}_{=1} \\ &= -2\rho_j + 2w_j \end{aligned}$$

Parte 2: Derivada parcial do termo de penalização L_1

► $L_1 = \lambda \sum_{j=0}^D |w_j|$

► $\lambda \frac{\partial}{\partial w_j} |w_j| = ???$



Parte 2: Subgradiente do termo de penalização L_1

$$\blacktriangleright L_1 = \lambda \sum_{j=0}^D |w_j|$$

$$\lambda \partial_{w_j} |w_j| = \begin{cases} -\lambda & \text{se } w_j < 0 \\ [-\lambda, \lambda] & \text{se } w_j = 0 \\ \lambda & \text{se } w_j > 0 \end{cases}$$

Juntando as partes

$$\text{RSS}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 = \sum_{i=1}^N \left(y_i - \sum_{j=0}^D w_j h_j(\mathbf{x}^{(i)}) \right)^2 + \lambda \sum_{j=0}^D |w_j|$$

Juntando as partes

$$\text{RSS}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 = \sum_{i=1}^N \left(y_i - \sum_{j=0}^D w_j h_j(\mathbf{x}^{(i)}) \right)^2 + \lambda \sum_{j=0}^D |w_j|$$

$$\lambda \partial_{w_j}(\text{custo lasso}) = \underbrace{2w_j - 2\rho_j}_{\text{do RSS}} + \underbrace{\begin{cases} -\lambda & \text{se } w_j < 0 \\ [-\lambda, \lambda] & \text{se } w_j = 0 \\ \lambda & \text{se } w_j > 0 \end{cases}}_{\text{da penalidade } L_1}$$

Juntando as partes

$$\text{RSS}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 = \sum_{i=1}^N \left(y_i - \sum_{j=0}^D w_j h_j(\mathbf{x}^{(i)}) \right)^2 + \lambda \sum_{j=0}^D |w_j|$$

$$\lambda \partial_{w_j}(\text{custo lasso}) = \underbrace{2w_j - 2\rho_j}_{\text{do RSS}} + \underbrace{\begin{cases} -\lambda & \text{se } w_j < 0 \\ [-\lambda, \lambda] & \text{se } w_j = 0 \\ \lambda & \text{se } w_j > 0 \end{cases}}_{\text{da penalidade } L_1}$$

$$= \begin{cases} 2w_j - 2\rho_j - \lambda & \text{se } w_j < 0 \\ [-2\rho_j - \lambda, -2\rho_j + \lambda] & \text{se } w_j = 0 \\ 2w_j - 2\rho_j + \lambda & \text{se } w_j > 0 \end{cases}$$

Solução Ótima: iguale subgradiente a zero

$$\partial_{w_j}(\text{custo lasso}) = \begin{cases} 2w_j - 2\rho_j - \lambda & \text{se } w_j < 0 \\ [-2\rho_j - \lambda, -2\rho_j + \lambda] & \text{se } w_j = 0 \\ 2w_j - 2\rho_j + \lambda & \text{se } w_j > 0 \end{cases}$$

Caso 1: $w_j < 0$

$$\begin{aligned} 2\hat{w}_j - 2\rho_j - \lambda &= 0 \\ \hat{w}_j &= \frac{2\rho_j + \lambda}{2} \\ &= \rho_j + \frac{\lambda}{2} \end{aligned}$$

Para $\hat{w}_j < 0$ precisamos que $\rho_j < -\frac{\lambda}{2}$

Solução Ótima: iguale subgradiente a zero

$$\partial_{w_j}(\text{custo lasso}) = \begin{cases} 2w_j - 2\rho_j - \lambda & \text{se } w_j < 0 \\ [-2\rho_j - \lambda, -2\rho_j + \lambda] & \text{se } w_j = 0 \\ 2w_j - 2\rho_j + \lambda & \text{se } w_j > 0 \end{cases}$$

Caso 2: $w_j = 0$: $[-2\rho_j - \lambda, -2\rho_j + \lambda]$ deve conter 0

$$-2\rho_j + \lambda \geq 0 \rightarrow \rho_j \leq \frac{\lambda}{2}$$

$$-2\rho_j - \lambda \leq 0 \rightarrow \rho_j \geq \frac{-\lambda}{2}$$

$$\frac{-\lambda}{2} \leq \rho_j \leq \frac{\lambda}{2}$$

Solução Ótima: iguale subgradiente a zero

$$\partial_{w_j}(\text{custo lasso}) = \begin{cases} 2w_j - 2\rho_j - \lambda & \text{se } w_j < 0 \\ [-2\rho_j - \lambda, -2\rho_j + \lambda] & \text{se } w_j = 0 \\ 2w_j - 2\rho_j + \lambda & \text{se } w_j > 0 \end{cases}$$

Caso 3: $w_j > 0$

$$\hat{w}_j = \rho_j - \frac{\lambda}{2}$$

Para $\hat{w}_j > 0$ precisamos que $\rho_j > \frac{\lambda}{2}$

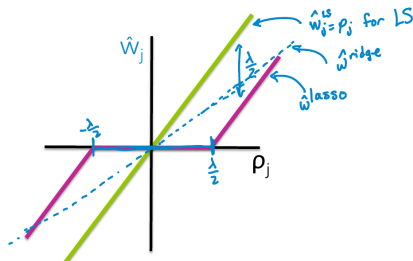
Solução Ótima: iguale subgradiente a zero

$$\partial_{w_j}(\text{custo lasso}) = \begin{cases} 2w_j - 2\rho_j - \lambda & \text{se } w_j < 0 \\ [-2\rho_j - \lambda, -2\rho_j + \lambda] & \text{se } w_j = 0 \\ 2w_j - 2\rho_j + \lambda & \text{se } w_j > 0 \end{cases}$$

$$\hat{w}_j = \begin{cases} \rho_j + \frac{\lambda}{2} & \text{se } \rho_j < -\frac{\lambda}{2} \\ 0 & \text{se } \rho_j \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right] \\ \rho_j - \frac{\lambda}{2} & \text{se } \rho_j > \frac{\lambda}{2} \end{cases}$$

Interpretação

$$\hat{w}_j = \begin{cases} \rho_j + \frac{\lambda}{2} & \text{se } \rho_j < -\frac{\lambda}{2} \\ 0 & \text{se } \rho_j \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right] \\ \rho_j - \frac{\lambda}{2} & \text{se } \rho_j > \frac{\lambda}{2} \end{cases}$$



Lasso com Coordinate Descent

Coordinate-Descent-Lasso(λ)

1 initialize $\hat{\mathbf{w}}$



2 **while** not converged

3 pick a coordinate j

4
$$\rho_j = \sum_{i=1}^N h_j(\mathbf{x}^{(i)})(y_i - \hat{y}_i(\hat{\mathbf{w}}_{-j}))$$

5
$$\hat{\mathbf{w}}_j = \begin{cases} \rho_j + \frac{\lambda}{2} & \text{se } \rho_j < -\frac{\lambda}{2} \\ 0 & \text{se } \rho_j \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right] \\ \rho_j - \frac{\lambda}{2} & \text{se } \rho_j > \frac{\lambda}{2} \end{cases}$$

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: 18/11/2016.]