Aula 4 - Regressão Não Linear e Equação Normal

João Florindo

Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas - Brasil
florindo@unicamp.br

Outline

Normalização/Padronização

- Restringir todo x_j ao mesmo intervalo de valores acelera o gradiente descendente.

NORMALIZAÇÃO (*scaling*):

$$x_j := \frac{x_j - \text{mínimo}(x_j)}{\text{máximo}(x_j) - \text{mínimo}(x_j)}$$

PADRONIZAÇÃO (*z-score*):

$$x_j := \frac{x_j - \text{média}(x_j)}{\text{desvio padrão}(x_j)}$$

Normalização/Padronização

- Restringir todo x_j ao mesmo intervalo de valores acelera o gradiente descendente.

NORMALIZAÇÃO (*scaling*):

$$x_j := \frac{x_j - \text{mínimo}(x_j)}{\text{máximo}(x_j) - \text{mínimo}(x_j)}$$

PADRONIZAÇÃO (*z-score*):

$$x_j := \frac{x_j - \text{média}(x_j)}{\text{desvio padrão}(x_j)}$$

Normalização/Padronização

- Restringir todo x_j ao mesmo intervalo de valores acelera o gradiente descendente.

NORMALIZAÇÃO (*scaling*):

$$x_j := \frac{x_j - \text{mínimo}(x_j)}{\text{máximo}(x_j) - \text{mínimo}(x_j)}$$

PADRONIZAÇÃO (*z-score*):

$$x_j := \frac{x_j - \text{média}(x_j)}{\text{desvio padrão}(x_j)}$$

Normalização/Padronização

Normalização *vs.* Padronização

- Normalização: atributos não seguem a distribuição normal.
- Padronização: atributos seguem a distribuição normal (não obrigatório). Preserva *outliers*.

Normalização/Padronização

Normalização *vs.* Padronização

- Normalização: atributos não seguem a distribuição normal.
- Padronização: atributos seguem a distribuição normal (não obrigatório). Preserva *outliers*.

Normalização/Padronização

Normalização *vs.* Padronização

- Normalização: atributos não seguem a distribuição normal.
- Padronização: atributos seguem a distribuição normal (não obrigatório). Preserva *outliers*.

Taxa de Aprendizado

- Inspecionar gráfico de $J(\theta)$ em função do número de iterações
- Parar, por exemplo, com 1000 iterações ou se $J(\theta) < 10^{-3}$
- $J(\theta)$ aumentando ou flutuando: reduzir α

Na regressão linear, $J(\theta)$ diminui em TODA iteração se α for pequeno o suficiente.

Testar vários valores de α. EXEMPLO:

$$\alpha = \cdots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \cdots$$

Taxa de Aprendizado

- Inspecionar gráfico de $J(\theta)$ em função do número de iterações
- Parar, por exemplo, com 1000 iterações ou se $J(\theta) < 10^{-3}$
- $J(\theta)$ aumentando ou flutuando: reduzir α

Na regressão linear, $J(\theta)$ diminui em TODA iteração se α for pequeno o suficiente.

Testar vários valores de α. EXEMPLO:

$$\alpha = \cdots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \cdots$$

Taxa de Aprendizado

- Inspecionar gráfico de $J(\theta)$ em função do número de iterações
- Parar, por exemplo, com 1000 iterações ou se $J(\theta) < 10^{-3}$
- $J(\theta)$ aumentando ou flutuando: reduzir α

Na regressão linear, $J(\theta)$ diminui em TODA iteração se α for pequeno o suficiente.

Testar vários valores de α. EXEMPLO:

$$\alpha = \cdots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \cdots$$

Taxa de Aprendizado

- Inspecionar gráfico de $J(\theta)$ em função do número de iterações
- Parar, por exemplo, com 1000 iterações ou se $J(\theta) < 10^{-3}$
- $J(\theta)$ aumentando ou flutuando: reduzir α

Na regressão linear, $J(\theta)$ diminui em TODA iteração se α for pequeno o suficiente.

Testar vários valores de α. EXEMPLO:

$$\alpha = \cdots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \cdots$$

Taxa de Aprendizado

- Inspecionar gráfico de $J(\theta)$ em função do número de iterações
- Parar, por exemplo, com 1000 iterações ou se $J(\theta) < 10^{-3}$
- $J(\theta)$ aumentando ou flutuando: reduzir α

Na regressão linear, $J(\theta)$ diminui em TODA iteração se α for pequeno o suficiente.

Testar vários valores de α. EXEMPLO:

$$\alpha = \cdots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \cdots$$

Outline

POSSÍVEIS MELHORIAS NA FUNÇÃO DE HIPÓTESE:

- Combinar atributos: em vez de usar altura e peso do paciente, usar IMC = peso/altura^2.
- Observar os exemplos de treinamento, se a relação entre atributos e saída for não linear, adicionar atributos não lineares.

EXEMPLOS

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2 + \theta_2 x_1^3$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 \sqrt{x_1}$$

POSSÍVEIS MELHORIAS NA FUNÇÃO DE HIPÓTESE:

- Combinar atributos: em vez de usar altura e peso do paciente, usar IMC = peso/altura^2.
- Observar os exemplos de treinamento, se a relação entre atributos e saída for não linear, adicionar atributos não lineares.

EXEMPLOS

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2 + \theta_2 x_1^3$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 \sqrt{x_1}$$

POSSÍVEIS MELHORIAS NA FUNÇÃO DE HIPÓTESE:

- Combinar atributos: em vez de usar altura e peso do paciente, usar IMC = peso/altura^2.
- Observar os exemplos de treinamento, se a relação entre atributos e saída for não linear, adicionar atributos não lineares.

EXEMPLOS

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2 + \theta_2 x_1^3$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 \sqrt{x_1}$$

- Algoritmo idêntico à regressão linear multivariada. Exemplo:

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2 + \theta_2 x_1^3.$$

  Definir novas variáveis

$$x_2 := x_1^2 \qquad x_3 := x_1^3.$$

- ATENÇÃO: Normalização é ainda mais necessária neste caso!

- Algoritmo idêntico à regressão linear multivariada. Exemplo:

$$h_\theta(x) = \theta_0 + \theta_1 x_1^2 + \theta_2 x_1^3.$$

Definir novas variáveis

$$x_2 := x_1^2 \qquad x_3 := x_1^3.$$

- ATENÇÃO: Normalização é ainda mais necessária neste caso!

Outline

- Minimização analítica de $J(\theta)$

Definimos a *matriz de design* **X**, o vetor de saídas **y** e o vetor de parâmetros θ:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(m)} \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

Então:

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

- Minimização analítica de $J(\theta)$

Definimos a *matriz de design* $\mathbf{X}$, o vetor de saídas $\mathbf{y}$ e o vetor de parâmetros θ:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(m)} \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

Então:

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

GRADIENTE DESCENDENTE *vs.* EQUAÇÃO NORMAL

Gradiente Descente	**Equação Normal**
Normalizar atributos	Não precisa normalizar
Precisa escolher α	Não precisa escolher α
Mais iterações	Sem iterações
$\mathcal{O}(kn^2)$	$\mathcal{O}(n^3)$ devido à inversa de X^TX
Funciona bem com n grande	Lento se n for grande

Comum que se use gradiente para $n > 10000$ e equação normal para n menor.

GRADIENTE DESCENDENTE *vs.* EQUAÇÃO NORMAL

Gradiente Descente	**Equação Normal**
Normalizar atributos	Não precisa normalizar
Precisa escolher α	Não precisa escolher α
Mais iterações	Sem iterações
$\mathcal{O}(kn^2)$	$\mathcal{O}(n^3)$ devido à inversa de X^TX
Funciona bem com n grande	Lento se n for grande

Comum que se use gradiente para $n > 10000$ e equação normal para n menor.

PROBLEMA: $\mathbf{X}^T\mathbf{X}$ pode não ser inversível.

SOLUÇÃO: Usar pseudo-inversa, função *pinv*.

CAUSAS:

- Atributos redundantes
- Atributos em excesso: $m < n$

PROBLEMA: $\mathbf{X}^T\mathbf{X}$ pode não ser inversível.

SOLUÇÃO: Usar pseudo-inversa, função *pinv*.

CAUSAS:

- Atributos redundantes
- Atributos em excesso: $m < n$

PROBLEMA: $\mathbf{X}^T\mathbf{X}$ pode não ser inversível.

SOLUÇÃO: Usar pseudo-inversa, função *pinv*.

CAUSAS:

- Atributos redundantes
- Atributos em excesso: $m < n$