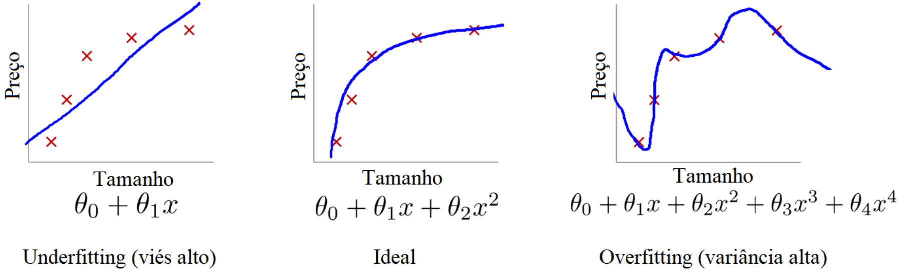
Aula 6 - *Overfitting* e Regularização

João Florindo

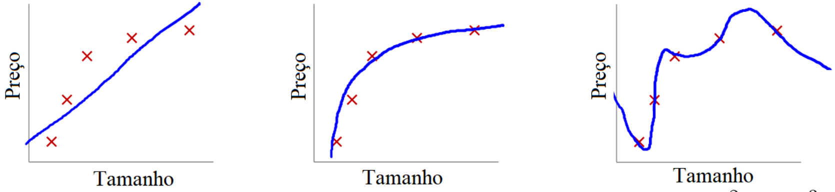
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas - Brasil
florindo@unicamp.br

Outline



Overfitting

Muitos atributos, a função de hipótese se ajusta muito bem ao treinamento, i.e., $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)}))^2 \approx 0$, mas não generaliza bem.



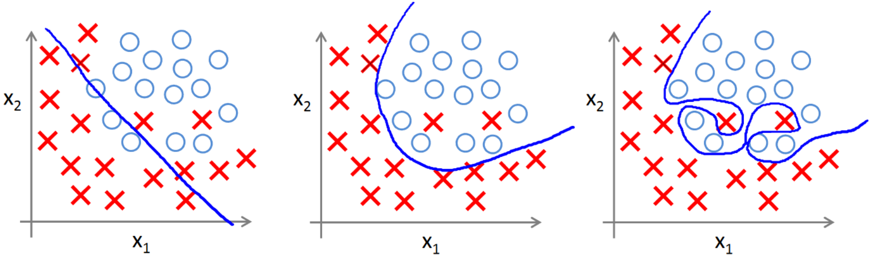
$\theta_0 + \theta_1 x$ | $\theta_0 + \theta_1 x + \theta_2 x^2$ | $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Underfitting (viés alto) | Ideal | Overfitting (variância alta)

Overfitting

Muitos atributos, a função de hipótese se ajusta muito bem ao treinamento, i.e., $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)}))^2 \approx 0$, mas não generaliza bem.

Na regressão logística:



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2$
$+\theta_3 x_1^2 + \theta_4 x_2^2$
$+\theta_5 x_1 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2$
$+\theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2$
$+\theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \ldots)$

Soluções

1. Reduzir o número de atributos
   - Manualmente
   - Algoritmo de seleção de atributos
2. Regularização
   - Mantém todos os atributos, mas reduz a magnitude (peso) dos parâmetros θ_j
   - Funciona bem quando temos muitos atributos e todos eles contribuem com algo (mesmo que pouco)

Soluções

1. Reduzir o número de atributos
   - Manualmente
   - Algoritmo de seleção de atributos
2. Regularização
   - Mantém todos os atributos, mas reduz a magnitude (peso) dos parâmetros θ_j
   - Funciona bem quando temos muitos atributos e todos eles contribuem com algo (mesmo que pouco)

Soluções

1. Reduzir o número de atributos
   - Manualmente
   - Algoritmo de seleção de atributos
2. Regularização
   - Mantém todos os atributos, mas reduz a magnitude (peso) dos parâmetros θ_j
   - Funciona bem quando temos muitos atributos e todos eles contribuem com algo (mesmo que pouco)

Soluções

1. Reduzir o número de atributos
   - Manualmente
   - Algoritmo de seleção de atributos
2. Regularização
   - Mantém todos os atributos, mas reduz a magnitude (peso) dos parâmetros θ_j
   - Funciona bem quando temos muitos atributos e todos eles contribuem com algo (mesmo que pouco)

Soluções

1. Reduzir o número de atributos
    - Manualmente
    - Algoritmo de seleção de atributos
2. Regularização
    - Mantém todos os atributos, mas reduz a magnitude (peso) dos parâmetros θ_j
    - Funciona bem quando temos muitos atributos e todos eles contribuem com algo (mesmo que pouco)

Soluções

1. Reduzir o número de atributos
   - Manualmente
   - Algoritmo de seleção de atributos
2. Regularização
   - Mantém todos os atributos, mas reduz a magnitude (peso) dos parâmetros θ_j
   - Funciona bem quando temos muitos atributos e todos eles contribuem com algo (mesmo que pouco)

Outline

INTUIÇÃO: Como forçar um polinômio de grau 4

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

a se aproximar de um de grau 2?

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000\theta_3^2 + 1000\theta_4^2.$$

Em geral:

$$\begin{cases} J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right] \\ \operatorname{argmin}_\theta J(\theta) \end{cases}$$

- Escolha de λ é uma decisão importante.

INTUIÇÃO: Como forçar um polinômio de grau 4

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

a se aproximar de um de grau 2?

$$\operatorname*{argmin}_{\theta} \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + 1000\theta_3^2 + 1000\theta_4^2.$$

Em geral:

$$\begin{cases} J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right] \\ \operatorname{argmin}_\theta J(\theta) \end{cases}$$

- Escolha de λ é uma decisão importante.

INTUIÇÃO: Como forçar um polinômio de grau 4

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

a se aproximar de um de grau 2?

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000\theta_3^2 + 1000\theta_4^2.$$

Em geral:

$$\begin{cases} J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right] \\ \operatorname{argmin}_\theta J(\theta) \end{cases}$$

- Escolha de λ é uma decisão importante.

INTUIÇÃO: Como forçar um polinômio de grau 4

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

a se aproximar de um de grau 2?

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000\theta_3^2 + 1000\theta_4^2.$$

Em geral:

$$\begin{cases} J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right] \\ \operatorname{argmin}_\theta J(\theta) \end{cases}$$

- Escolha de λ é uma decisão importante.

Regressão Linear Regularizada

Repita até convergir: $\{$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} + \frac{\lambda}{m} \theta_j \right], \quad j = 1, \cdots n$$

$\}$

O caso para $j = 1, \cdots n$ pode ser reescrito como:

$$\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m}\right) - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}.$$

Regressão Linear Regularizada

Repita até convergir: $\{$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} + \frac{\lambda}{m} \theta_j \right], \quad j = 1, \cdots n$$

$\}$

O caso para $j = 1, \cdots n$ pode ser reescrito como:

$$\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m} \right) - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}.$$

Equação Normal Regularizada

Definindo-se:

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix} \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

No caso regularizado, θ é obtido por

$$\theta = (X^T X + \lambda L)^{-1} X^T y,$$

em que L é a matriz $(n+1) \times (n+1)$ dada por:

$$L = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Equação Normal Regularizada

Definindo-se:

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix} \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

No caso regularizado, θ é obtido por

$$\theta = (X^T X + \lambda L)^{-1} X^T y,$$

em que L é a matriz $(n+1) \times (n+1)$ dada por:

$$L = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Regressão Logística Regularizada

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2.$$

Gradiente:

Repita até convergir: $\{$

$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_0^{(i)}$

$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_j^{(i)} + \frac{\lambda}{m}\theta_j\right], j = 1, \cdots, n$

$\}$

- Lembre-se que neste caso:

$$h_\theta(x) = \frac{1}{1+\mathrm{e}^{-\theta^T x}}.$$

Regressão Logística Regularizada

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2.$$

Gradiente:

Repita até convergir: $\{$

$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_0^{(i)}$

$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_j^{(i)} + \frac{\lambda}{m}\theta_j\right], j = 1, \cdots, n$

$\}$

- Lembre-se que neste caso:

$$h_\theta(x) = \frac{1}{1+\mathrm{e}^{-\theta^T x}}.$$

Regressão Logística Regularizada

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2.$$

Gradiente:

$$\boxed{\begin{array}{l}\text{Repita até convergir: } \{\\ \theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_0^{(i)}\\ \theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_j^{(i)} + \frac{\lambda}{m}\theta_j\right], j = 1,\cdots,n\\ \}\end{array}}$$

- Lembre-se que neste caso:

$$h_\theta(x) = \frac{1}{1+\mathrm{e}^{-\theta^T x}}.$$