

Modelos Lineares

Otaviano da Cruz Neto

Instituto de Ciencias Exatas - ICEx / UFF

30/05/2018

Introdução

- ▶ **O que é o método de Regressão?**

O método de Regressão é uma ferramenta que leva em consideração a dependência entre as variáveis que caracterizam os dados . Ou seja, a cada valor de entrada é associado a um valor dado por uma Função Target, o padrão a ser aprendido.

- ▶ **Tipos de Regressão**

Os dois tipos de regressão que serão exibidos serão a Regressão Linear e a Regressão Logística. A regressão linear admite a dependência linear entre as variáveis e busca um hiperplano que melhor aproxima a configuração. Já a regressão logística utiliza a associação de cada conjunto de características a respostas binárias, por exemplo, 0 ou 1.

Introdução

► **Dados Individuais** ($X^{(i)}$ e $Y^{(i)}$)

$$X^{(i)} = [x_1, x_2, \dots, x_m] \quad (1)$$

$$Y^{(i)} = y^{(i)} \quad (2)$$

► **Dados da Amostra** ($X^{(i)}, Y^{(i)}$)

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_m^{(1)} \\ 1 & x_1^{(2)} & \dots & x_m^{(2)} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(N)} & \dots & x_m^{(N)} \end{bmatrix} \quad (3)$$

Introdução

$$Y = \begin{bmatrix} y^{(0)} \\ y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix} \quad (4)$$

Regressão Linear

- ▶ **Vetor Normal à Hipótese**

$$W = [w_0, w_1, w_2, \dots, w_m] \quad (5)$$

- ▶ **Hipótese Linear**

$$h(w) = XW^T \quad (6)$$

- ▶ **Caracterização dos Erros dentro (E_{in}) da amostra e fora dela (E_{out})**

A preocupação com o erro que envolve o aprendizado de um padrão é evidente já que há a necessidade de aplicar em outros conjuntos fora da amostragem. Neste caso temos,

$$E_{in}(W) = \frac{1}{N} \sum_{n=1}^N \left(X^{(n)} W^T - Y^{(n)} \right)^2 = \frac{1}{N} \left(XW^T - Y \right)^2 \quad (7)$$

Regressão Linear

- **Gradiente de E_{in}** ($\vec{\nabla} E_{in}$)

$$\vec{\nabla} E_{in} = \frac{2}{N} X^T (XW^T - Y) \quad (8)$$

- **Gradiente decrescente (Solução Numérica)** : O método do gradiente decrescente é uma ferramenta que utiliza a propriedade do operador gradiente de apontar sempre na direção de máximo crescimento, afim de minimizar o erro dentro da amostra.

$$W_{t+1}^{\rightarrow} = W_t^{\rightarrow} - \alpha \vec{\nabla} E_{in} \quad (9)$$

Regressão Linear

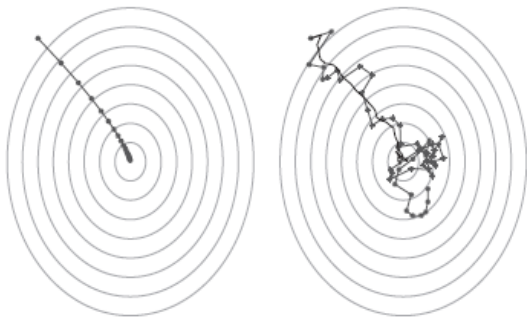


Figura 1: Diferença de valores de passos e suas convergência.

Solução da Regressão Linear

► Normalização (Solução analítica)

$$\vec{\nabla} E_{in} = \frac{2}{N} X^T (XW^T - Y) = 0 \quad (10)$$

$$W^T = X^\dagger Y \quad (11)$$

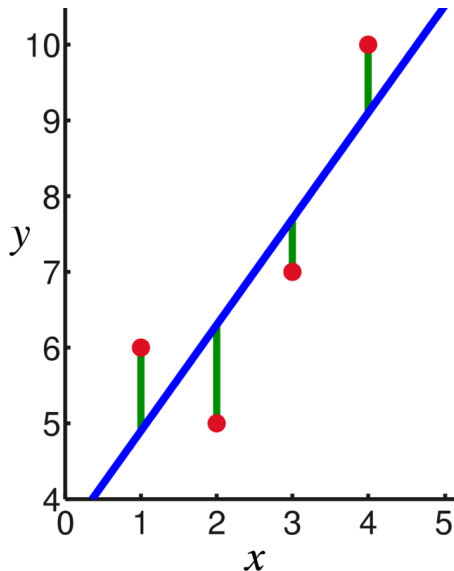
$$X^\dagger = (X^T X)^{-1} X^T \quad (12)$$

► Regressão Linear e PLA(Hipótese Inicial)

Muitas vezes na aplicação do Perceptron Learning Algorithm é inicialmente estipulado os valores do vetor normal (W) de maneira aleatória, e , por isso, pode haver uma demora em relação à convergência do algoritmo. Assim, afim de evitar tal situação, iniciar com uma hipótese que caracteriza melhor a amostragem é uma boa estratégia para diminuir o tempo de convergência. Ou seja, aplicando o algoritmo de regressão linear na amostragem e, a partir do padrão da regressão, aplicar o PLA.

Regressão Linear

► Visualização Gráfica



Interpretação Probabilística

- **Gaussiana (IID - Independente e identicamente distribuídos)**

Tomando que

$$Y_i = X_i W^T + \epsilon_i \quad (13)$$

Podemos expressar $\epsilon_i \approx N(0, \sigma^2)$, então a densidade de ϵ_i é dada por:

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - XW^T)^2}{2\sigma^2}\right) \quad (14)$$

Então a probabilidade(Likelihood) da Hipótese Linear é:

$$L(W) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - XW^T)^2}{2\sigma^2}\right) \quad (15)$$

Interpretação Probabilística

É fundamental diferenciar os papéis das duas expressões (14 e 15), respectivamente. Na equação 14 tem-se a probabilidade individual dentro da amostra em função de Y^i fixando W . Já na 15 temos a probabilidade geral da amostra em função de W , pois é justamente o valor de W que determinará o erro da amostra como um todo.

$$L(W) = L(W; X, Y) = p(Y|X; W) \quad (16)$$

Então derivando o $\ln L(w)$ temos:

$$l(W) = \ln L(w) \quad (17)$$

$$= \ln \prod_{i=1}^N \frac{2}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - XW^T)^2}{2\sigma^2}\right) \quad (18)$$

Interpretação Probabilística

$$= \sum_{i=1}^N \ln \frac{2}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - XW^T)^2}{2\sigma^2}\right) \quad (19)$$

$$= N \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^N (Y_i - X_i W^T)^2 \quad (20)$$

Maximizando temos:

$$l(W) = \frac{1}{2} \sum_{i=1}^N (Y_i - X_i W^T)^2 \quad (21)$$

Aplicação

► Dados

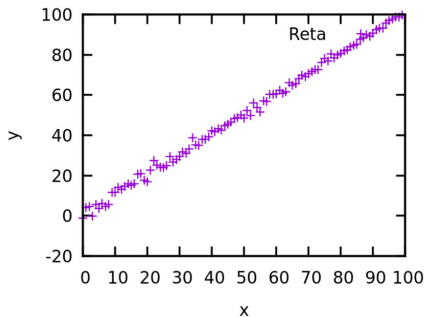


Figura 2: Dados criados a partir da reta $X=Y$.

Aplicação

► Função Custo a cada Iteração

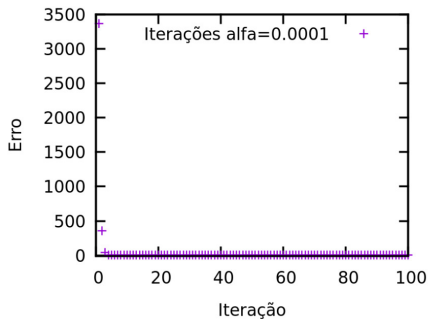


Figura 3: Gráfico de Custo por quantidade de iterações.

Aplicação

► Resultado Final

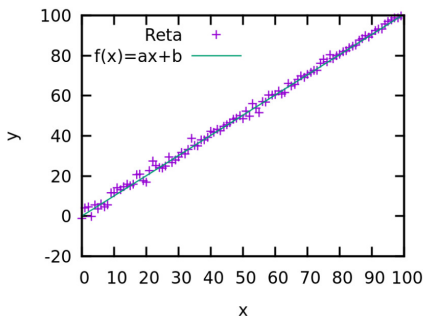


Figura 4: Gráfico da Reta obtida analiticamente de coeficiente angular $a = 1.00$ e coeficiente linear $b = -0.6$ criada a partir dos dados gerados.

Aplicação

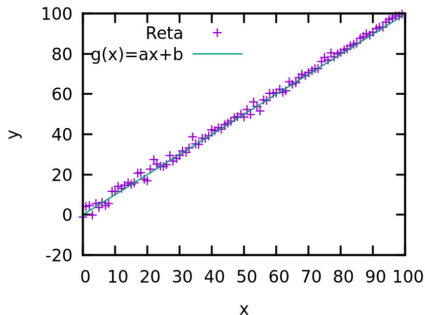


Figura 5: Gráfico da Reta obtida numericamente de coeficiente angular $a=0.99$ e coeficiente linear $b = 0.01$ criada a partir dos dados gerados.

Regressão Logística

► Classificação

A classificação é um problema de regressão que associa a cada valor da amostra um valor discreto. Neste caso vamos analisar para classificação binária (Positivo ou Negativo, 0 ou 1, -1 ou 1, etc).

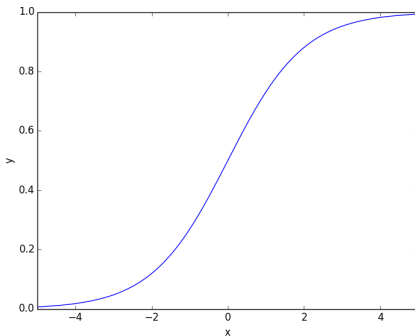
► Regressão Logística

O método de regressão logística utiliza a probabilidade de um determinado dado da amostra ser classificado por dos um valores discretos. A hipótese da regressão logística é que a probabilidade pode ser descrita como uma função logística (Sigmoid Function) que é dada por:

$$h_w(X^{(i)}) = \frac{1}{1 + e^{-X^{(i)} W^T}} \quad (22)$$

Regressão Logística

► Sigmoid Function



O comportamento da função é tal que a função tende a 1 quando $x \rightarrow \infty$ e quando $x \rightarrow -\infty$ a função tende a zero.

Regressão Logística

► Método do Gradiente Decrescente

Para utilizar essa técnica já descrita anteriormente é necessário o cálculo da derivada de $h_W(X^{(i)})$. Então, $h'(W)$ é dada por:

$$\begin{aligned}\frac{d}{dW} h_W(X^{(i)}) &= \frac{d}{dW} \frac{1}{1 + e^{-X^{(i)} W^T}} = \frac{1}{1 + e^{-X^{(i)} W^T}} (e^{-X^{(i)} W^T}) \\ &= \frac{1}{1 + e^{-X^{(i)} W^T}} \left(1 - \frac{1}{1 + e^{X^{(i)} W^T}} \right) = h_W(X^{(i)})(1 - h_W(X^{(i)}))\end{aligned}\tag{23}$$

Assim, de maneira semelhante ao modelo de Regressão Linear em que é adotado um conjunto de superposições de vetores derivadas, aqui é possível utilizar superposições probabilísticas a fim de maximizar o likelihood. Assumindo que a probabilidade de dado um $X^{(i)}$ e um vetor normal à hipótese ser classificado com o valor 1 é $h(W)$ e que a probabilidade dos mesmos vetor normal e $X^{(i)}$ ser classificado com o valor 0 é $1 - h(W)$.

Regressão Logística

- Likelihood

Ou seja,

$$P(Y^{(i)} = 1|X^{(i)}, W) = h_W(X^{(i)})$$

$$P(Y^{(i)} = 0|X^{(i)}, W) = 1 - h_W(X^{(i)})$$

Essa configuração de probabilidade condicional pode ser escrita de maneira mais funcional como:

$$p(Y^{(i)}|X^{(i)}, W) = (h_W(X^{(i)}))^{Y^{(i)}}(1 - h_W(X^{(i)}))^{1-Y^{(i)}} \quad (24)$$

Assim, a probabilidade para um número N de amostras IID é dada pela multiplicação de todas as probabilidades individuais.

Regressão Logística

- Likelihood

Ou seja:

$$L(W) = \prod_{i=1}^N (h_W(X^{(i)}))^{Y^{(i)}} (1 - h_W(X^{(i)}))^{1-Y^{(i)}} \quad (25)$$

Assim como na interpretação probabilística da Regressão Linear, é necessário maximizar o $\log(L(W))$:

$$\sum_{i=1}^N Y^{(i)} \log h_W(X^{(i)}) + (1 - Y^{(i)}) \log(1 - h_W(X^{(i)})) \quad (26)$$

Então, derivando:

$$\frac{\partial}{\partial W} l(W) = \left(\frac{y}{h_W(X^{(i)})} - (1 - y) \frac{1}{1 - h_W(X^{(i)})} \right) \frac{\partial}{\partial W} h_W(X^{(i)}) \quad (27)$$

Regressão Logística

- Likelihood

$$\frac{\partial}{\partial W} l(W) = (Y^{(i)} - h_W(X^{(i)}))X^{(i)} \quad (28)$$

Assim, o método do gradiente crescente:

$$W_{t+1} = W_t + \alpha(Y^{(i)} - h_W(X^{(i)}))X^{(i)} \quad (29)$$

Resultados

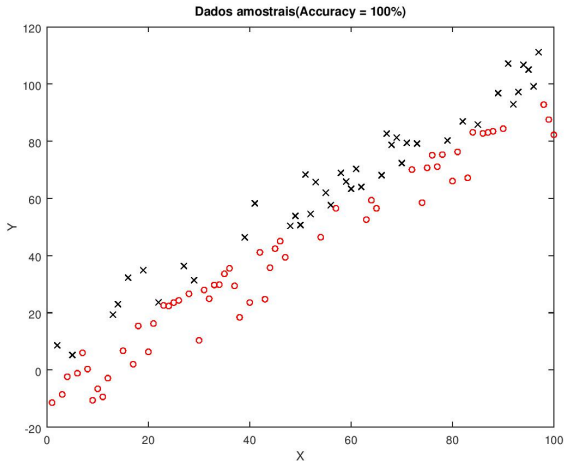


Figura 6: Dados Amostrais classificados por valores binário(0,1), acima da reta é classificado por 1 e abaixo por 0.

Resultados

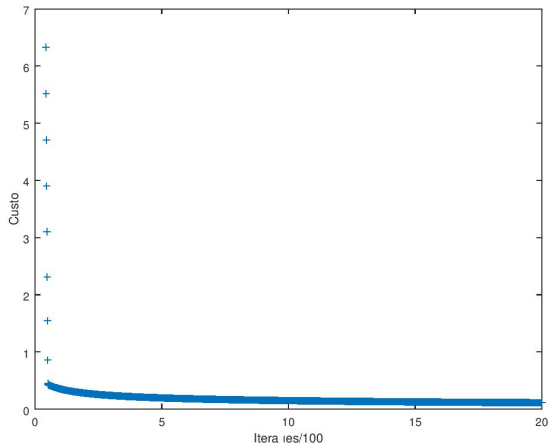


Figura 7: Valores do custo em função do número de iteração.

Resultados

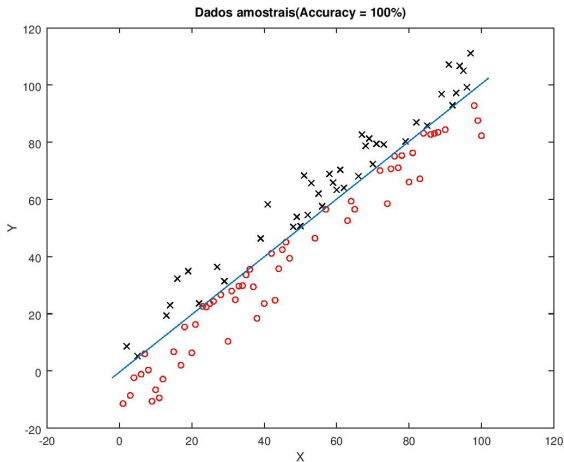


Figura 8: Amostra e Reta de decisão obtida após a implementação.

Referências

- [1] <http://www.portalaction.com.br/analise-de-regressao>. Acessado em 04/05/2018.
- [2] https://www.researchgate.net/figure/A-plot-of-the-gradient-descent-algorithm-left-and-the-stochastic-gradient-descent_fig1_303257470. Acessado em 04/05/2018.