Machine Learning

Dia 2 - Modelos Lineares

ImageU - Grupo de Pesquisa em Machine Learning e Visão Computacional
https://imageu.github.io/

Curso de Verão 2022

Instituto de Matemática e Estatística - IME USP

Programa

1. Representação de Dados
2. Regressão Linear
3. Classificação com Regressão Linear

Representação de Dados

Representação de Dados

- Já vimos que os dados são representados como elementos do $\mathbb{R}^d$:
    - Imagem $\longrightarrow \mathbf{x} = (x_1, x_2, \ldots, x_d)$
    - Paciente $\longrightarrow \mathbf{x} = (x_1, x_2, \ldots, x_d)$
    - Vídeo $\longrightarrow \mathbf{x} = (x_1, x_2, \ldots, x_d)$
- $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ é chamado de *feature vector*

Representação de dados

- Cedo ou tarde, na prática enfrentaremos alguns desafios
  - Informações faltantes
  - Valores com escalas distintas (cm, km)
  - Valores categóricos
  - Valores ruidosos
- Trataremos desses problemas mais adiante

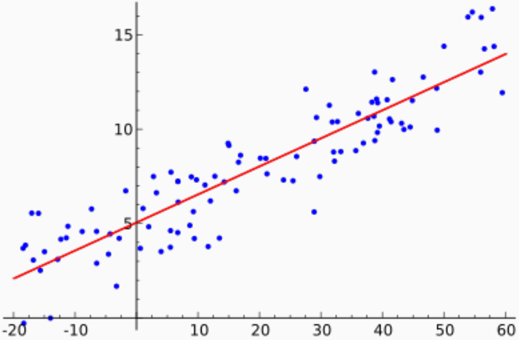
Regressão Linear

Regressão Linear

- Alvo é uma função $f : X \to Y$ na qual $Y \in \mathbb{R}$ (contínuo)
- Exemplos:
  - Estimar a temperatura da superfície via imagens de satélite
  - Estimar o preço de uma ação
  - Estimar o tempo de viagem de A a B
- Regressão linear faz sentido se há uma relação aproximadamente linear entre X e Y

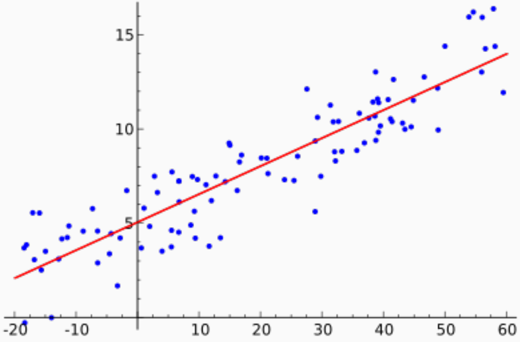
Regressão Linear

- Os pontos azuis $(x^{(n)}, y^{(n)})$ são os exemplos de treinamento
- Há uma relação linear entre x e y
- Família de hipóteses adequada: $g(x) = \hat{w}_0 + \hat{w}_1 \, x$
- Ou, em notação vetorial, $g(\mathbf{x}) = \widehat{\mathbf{w}}^T \mathbf{x}$, $\widehat{\mathbf{w}} = [\hat{w}_0, \hat{w}_1]$ e $\mathbf{x} = [1, x]$



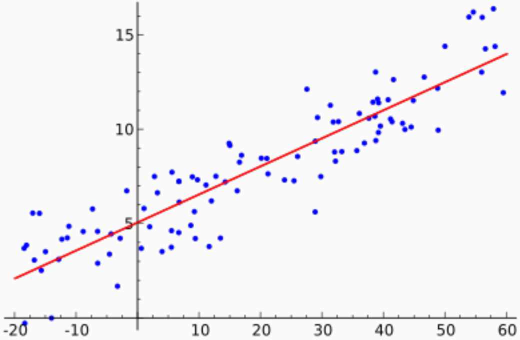
Como encontrar $\widehat{\mathbf{w}}$?

- Como encontrar os valores de $\widehat{\mathbf{w}}$?
  - Solução analítica: $\widehat{\mathbf{w}} = (X^T X)^{-1} X^T y$
  - Método iterativo



Como encontrar $\widehat{\mathrm{w}}$?

- Quando usamos a solução analítica? Na prática, quase nunca:
  - Inverter a matriz $(X^TX)_{(d+1)\times(d+1)}$ tem custo $O(n^3)$
  - Calcular X^TX também é caro (N pode ser muito grande)



Como encontrar $\widehat{w}$?

- Solução: resolver o problema iterativamente com *gradient descent* (descida do gradiente, método do gradiente, gradiente descendente)
  - Função de custo: $J(w_0, w_1) = \frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)} - \hat{y}^{(n)}\right)^2$
  - Em que $\hat{y}^{(n)} = g(x^{(n)}) = w_0 + w_1\, x^{(n)}$
  - Ou em notação vetorial: $J(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \mathbf{w}^T\mathbf{x}^{(n)})^2$
  - Objetivo é encontrar **w** que minimize o custo $J(\mathbf{w})$
- Mas antes, como saímos de $\widehat{\mathbf{w}} = (X^TX)^{-1}X^Ty$ e chegamos nessa função de custo?

Derivação da Equação Vetorial

- No caso geral:
- $\mathbf{x^{(n)}} = (1, x_1^{(n)}, x_2^{(n)}, \dots, x_d^{(n)}) \in \{1\} \times \mathbb{R}^d$
- $x_0^{(n)} = 1$
- $\mathbf{w} = (w_0, w_1, w_1, \dots, w_d) \in \mathbb{R}^{d+1}$
- $h_{\mathbf{w}}(\mathbf{x^{(n)}}) = \sum_{i=0}^{d} w_i \, x_i^{(n)} = \begin{bmatrix} w_0 & w_1 & \dots & w_d \end{bmatrix} \begin{bmatrix} 1 \\ x_1^{(n)} \\ \dots \\ x_d^{(n)} \end{bmatrix} = \mathbf{w}^T \, \mathbf{x^{(n)}}$
- $J(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)} \right)^2$

Derivação da Equação Vetorial

- Função de custo: $J(\mathbf{w}) = \dfrac{1}{N}\sum_{n=1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)^2$

Considere o vetor das diferenças (resíduos):

$$\begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) - y^{(1)} \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) - y^{(2)} \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) - y^{(N)} \end{bmatrix}$$

Derivação da Equação Vetorial

- Função de custo: $J(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)^2$

Considere o vetor das diferenças (resíduos):

$$\begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) - y^{(1)} \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) - y^{(2)} \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) - y^{(N)} \end{bmatrix} = \begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) \end{bmatrix} - \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}}_{\mathbf{y}}$$

Derivação da Equação Vetorial

- Função de custo: $J(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)^2$

Considere o vetor das diferenças (resíduos):

$$\begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) - y^{(1)} \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) - y^{(2)} \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) - y^{(N)} \end{bmatrix} = \begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) \end{bmatrix} - \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}}_{\mathbf{y}} = \begin{bmatrix} \mathbf{w}^T\mathbf{x}^{(1)} \\ \mathbf{w}^T\mathbf{x}^{(2)} \\ \vdots \\ \mathbf{w}^T\mathbf{x}^{(N)} \end{bmatrix} - \mathbf{y}$$

Derivação da Equação Vetorial

- Função de custo: $J(\mathbf{w}) = \dfrac{1}{N}\sum_{n=1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)^2$

Considere o vetor das diferenças (resíduos):

$$\begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) - y^{(1)} \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) - y^{(2)} \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) - y^{(N)} \end{bmatrix} = \begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) \end{bmatrix} - \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}}_{\mathbf{y}} = \begin{bmatrix} \mathbf{w}^T\mathbf{x}^{(1)} \\ \mathbf{w}^T\mathbf{x}^{(2)} \\ \vdots \\ \mathbf{w}^T\mathbf{x}^{(N)} \end{bmatrix} - \mathbf{y} =$$

$$\begin{bmatrix} w_0 + w_1 x_1^{(1)} + \ldots + w_d x_d^{(1)} \\ w_0 + w_1 x_1^{(2)} + \ldots + w_d x_d^{(2)} \\ \vdots \\ w_0 + w_1 x_1^{(N)} + \ldots + w_d x_d^{(N)} \end{bmatrix} - \mathbf{y}$$

Derivação da Equação Vetorial

- Função de custo: $J(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(h_{\mathbf{w}}(\mathbf{x}^{(n)}) - y^{(n)}\right)^2$

Considere o vetor das diferenças (resíduos):

$$\begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) - y^{(1)} \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) - y^{(2)} \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) - y^{(N)} \end{bmatrix} = \begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}^{(1)}) \\ h_{\mathbf{w}}(\mathbf{x}^{(2)}) \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}^{(N)}) \end{bmatrix} - \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}}_{\mathbf{y}} = \begin{bmatrix} \mathbf{w}^T\mathbf{x}^{(1)} \\ \mathbf{w}^T\mathbf{x}^{(2)} \\ \vdots \\ \mathbf{w}^T\mathbf{x}^{(N)} \end{bmatrix} - \mathbf{y} =$$

$$\begin{bmatrix} w_0 + w_1 x_1^{(1)} + \ldots + w_d x_d^{(1)} \\ w_0 + w_1 x_1^{(2)} + \ldots + w_d x_d^{(2)} \\ \vdots \\ w_0 + w_1 x_1^{(N)} + \ldots + w_d x_d^{(N)} \end{bmatrix} - \mathbf{y} = \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & \ldots & x_d^{(1)} \\ 1 & x_1^{(2)} & \ldots & x_d^{(2)} \\ \vdots & & & \\ 1 & x_1^{(N)} & \ldots & x_d^{(N)} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_N \end{bmatrix} - \mathbf{y}}_{\mathbf{Xw}-\mathbf{y}}$$

Derivação da Equação Vetorial

O vetor de resíduos pode ser escrito

$$\begin{bmatrix} h_{\mathbf{w}}(\mathbf{x}_1) - y_1 \\ h_{\mathbf{w}}(\mathbf{x}_2) - y_2 \\ \vdots \\ h_{\mathbf{w}}(\mathbf{x}_N) - y_N \end{bmatrix} = \mathbf{Xw} - \mathbf{y}$$

Precisamos dos resíduos ao quadrado:

$$\begin{bmatrix} (h_{\mathbf{w}}(\mathbf{x}_1) - y_1)^2 \\ (h_{\mathbf{w}}(\mathbf{x}_2) - y_2)^2 \\ \vdots \\ (h_{\mathbf{w}}(\mathbf{x}_N) - y_N)^2 \end{bmatrix} = (\mathbf{Xw} - \mathbf{y})^T(\mathbf{Xw} - \mathbf{y})$$

que é equivalente a $||\mathbf{Xw} - \mathbf{y}||^2$

Derivação da Equação Vetorial

Para minimizar

$$E(\mathbf{w}) = \frac{1}{N}||\mathbf{X}\mathbf{w} - \mathbf{y}||^2$$

Encontramos o ponto de mínimo via gradiente

$$\nabla E(\mathbf{w}) = \frac{2}{N}\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y}) = 0$$

que é equivalente à nossa solução analítica

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

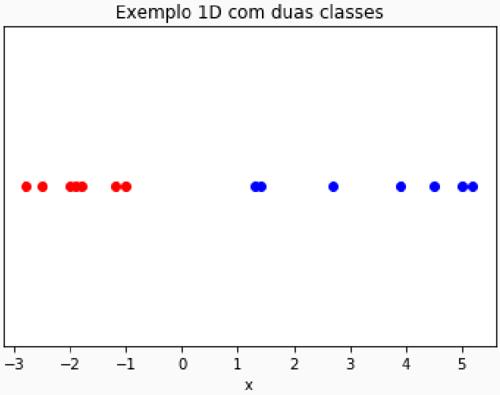
Classificação com Regressão Linear

Classificação com regressão linear

- Regressão linear aprende uma função real $y = f(x) \in \mathbb{R}$
- Funções binárias também podem ser vistas como reais, afinal $\pm 1 \in \mathbb{R}$
- Vamos tentar usar regressão linear para encontrar $\mathbf{w}$ tal que $\mathbf{w}^T \mathbf{x}_n \approx y_n = \pm 1$
- Nesse caso, $sinal(\mathbf{w}^T \mathbf{x}_n)$ deve se aproximar de $y_n = \pm 1$

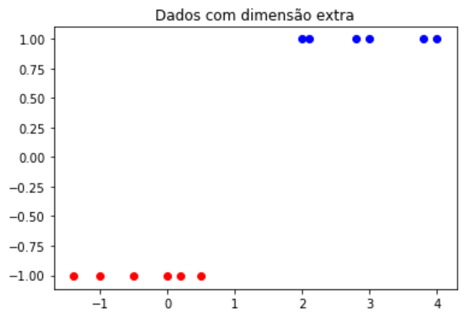
Classificação com regressão linear

- Um exemplo simples:
  - se $x < 1$ então vermelho
  - se $x \geq 1$ então azul
- Mas como encontrar essa fronteira ?



Classificação com regressão linear

- Podemos adotar
  - $y = -1$: exemplo negativo
  - $y = 1$: exemplo positivo
- e aplicar a regressão linear!
- Uma decisão simples seria $g(x) = sinal(w^T x + b)$!

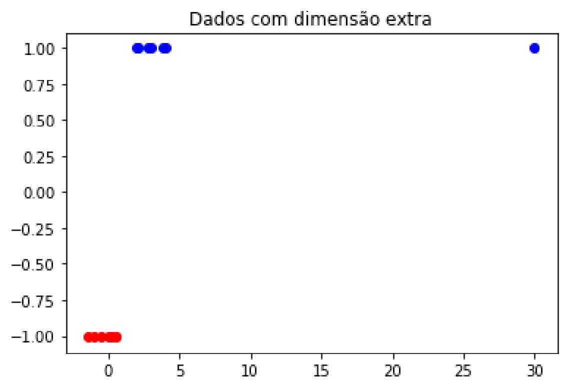


Classificação com regressão linear

- Podemos adotar
  - $y = -1$: exemplo negativo
  - $y = 1$: exemplo positivo
- e aplicar a regressão linear!
- Fronteira de decisão $g(x) = sinal(w^T x + b) = 0$ em verde.

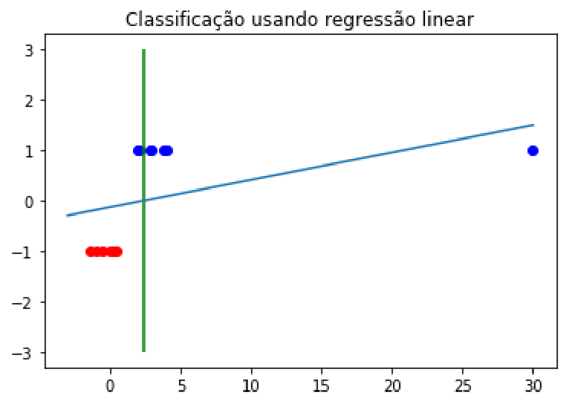
Classificação com regressão linear

- E neste exemplo, a regressão linear daria certo?



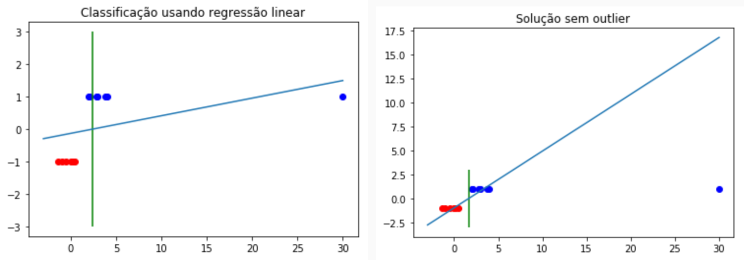
Classificação com regressão linear

- Não muito:



Classificação com regressão linear

- Por que não escolher a solução da direita?
- Estamos otimizando $||Xw - y||^2 = \sum_{i=1}^{N}(w^T x_i - y_i)^2$!



- Fazer classificação com regressão linear em geral não funciona.
- O modelo linear para classificação é chamado Regressão Logística

Próxima Aula: Regressão Logística e Gradient Descent