Classificação x Regressão

$\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)}) \in X \times Y, n = 1, \ldots, N\}$

We would like to "learn" $f : X \to Y \quad (\ X = \mathbb{R}^d\ )$

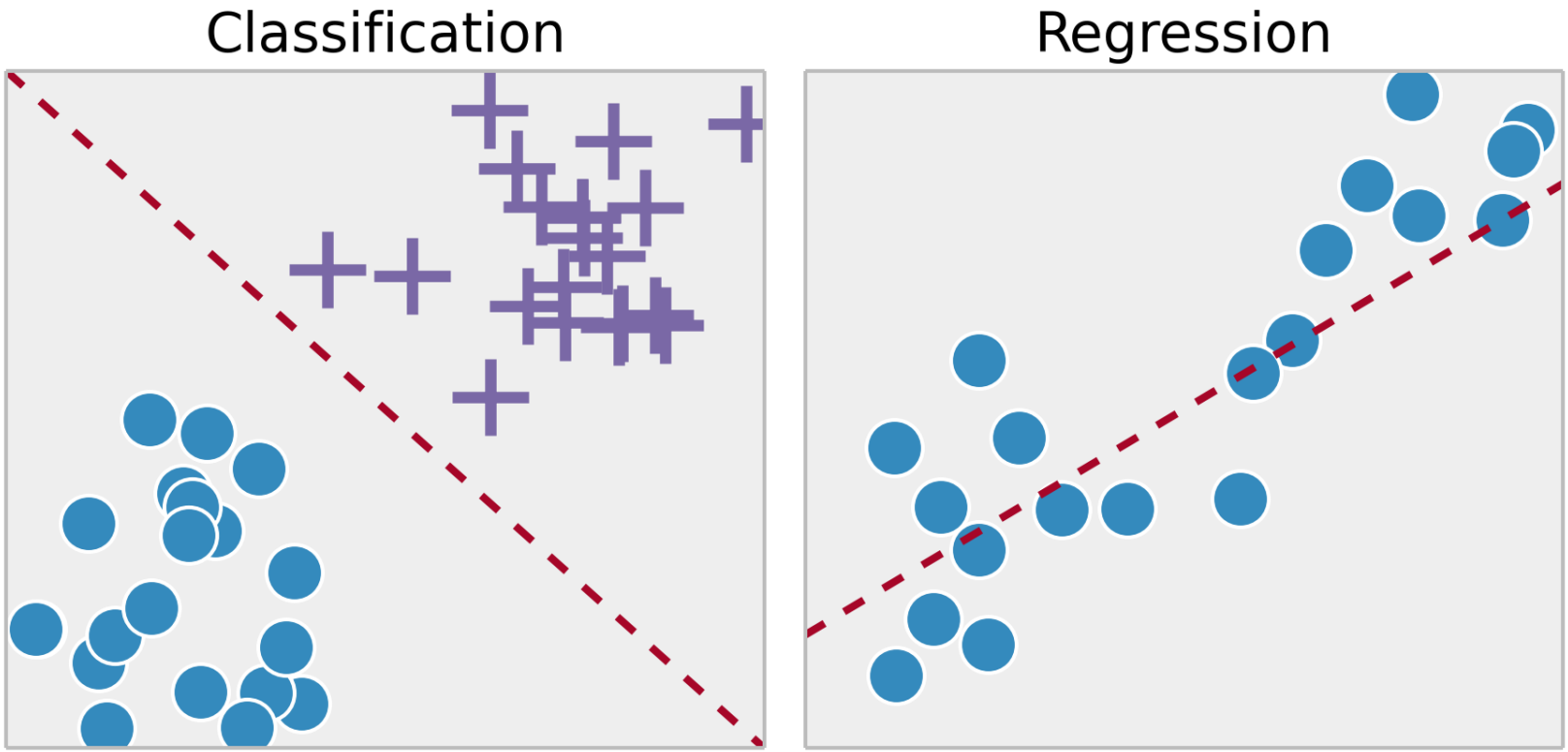
Classificação

y é um inteiro que representa uma classe (ou categoria)

Regressão

y é um número real

Classificação x Regressão

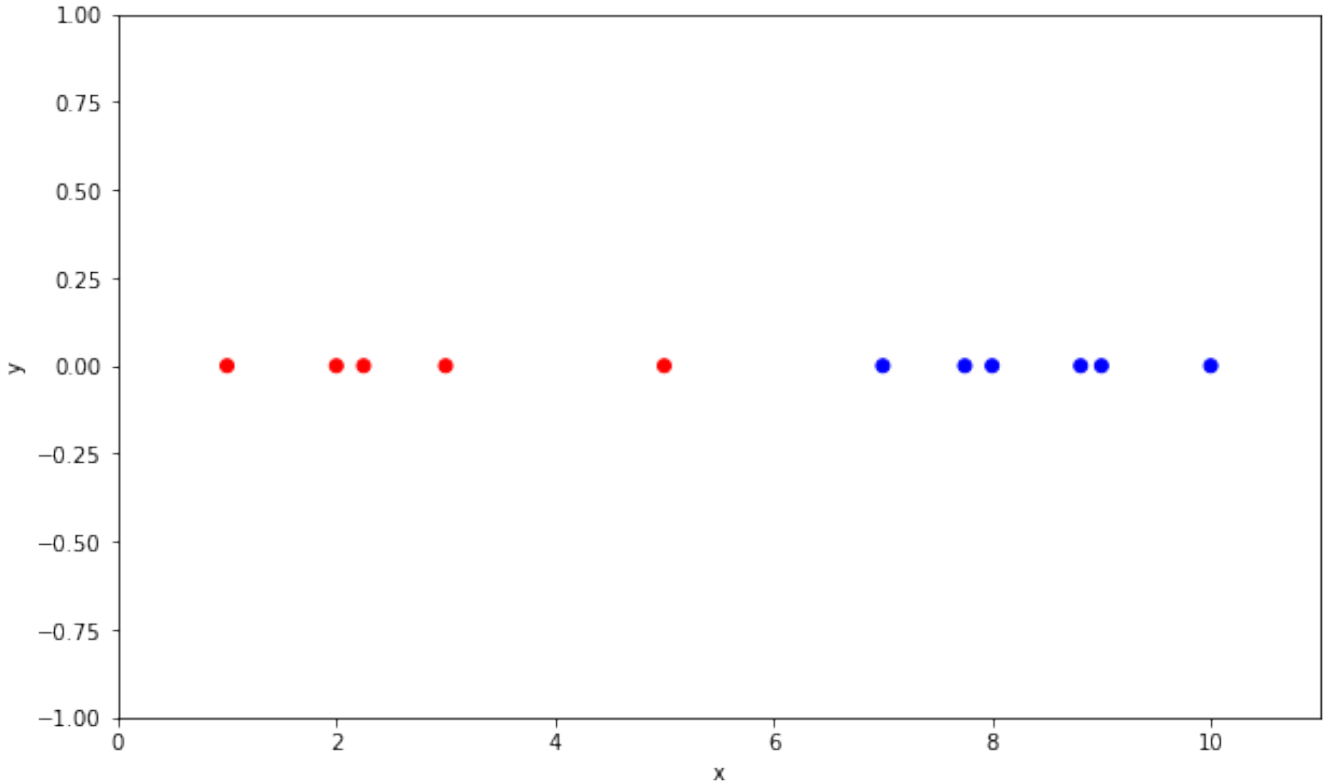


In the example above:

- Classification: $\mathbf{x} \in \mathbb{R}^2$, two classes
- Regression: $\mathbf{x} \in \mathbb{R}$, $y \in \mathbb{R}$

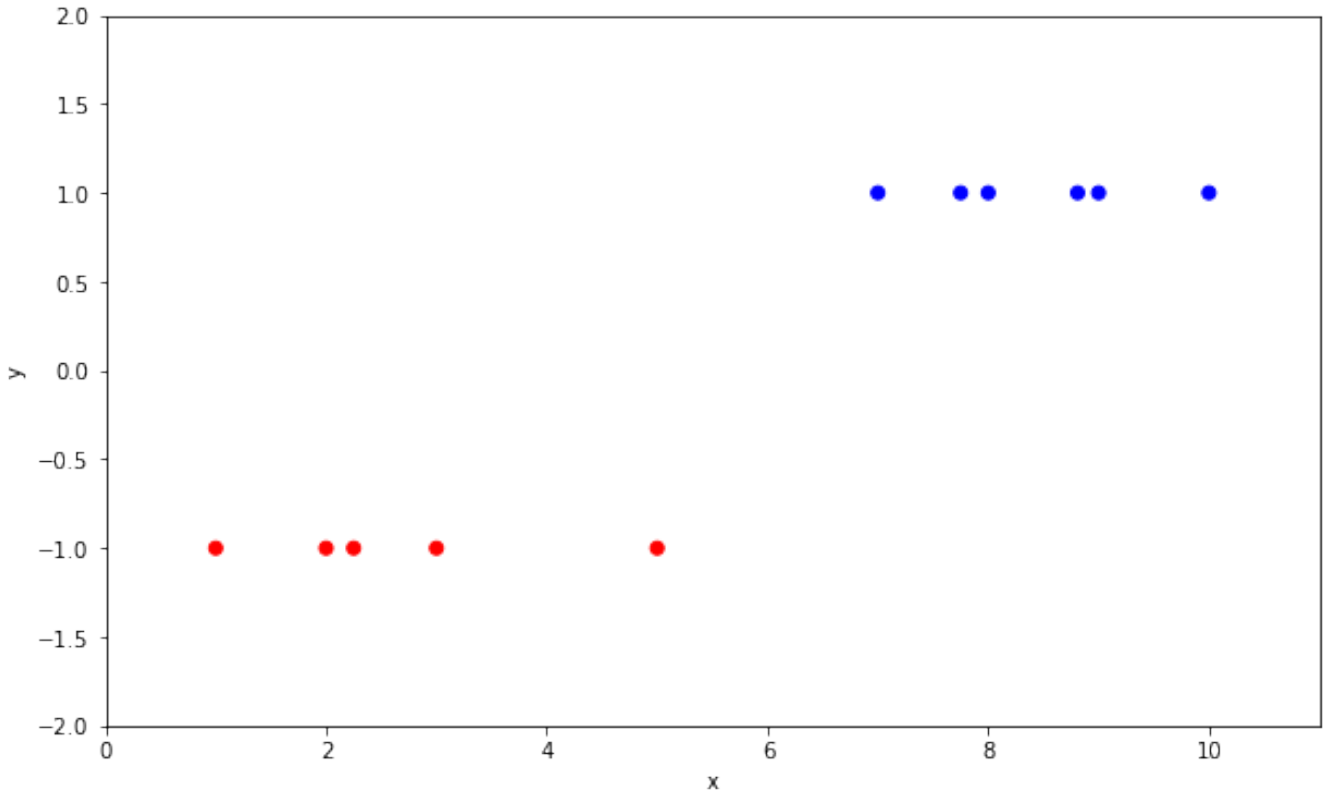
Classificação binária via regressão linear

Example: $D_X = \{1, 2, 2.25, 3, 5, 7, 7.75, 8, 8.81, 9, 10\}$



Classificação binária via regressão linear

Amostras negativas: $y = -1$
Amostras positivas; $y = +1$



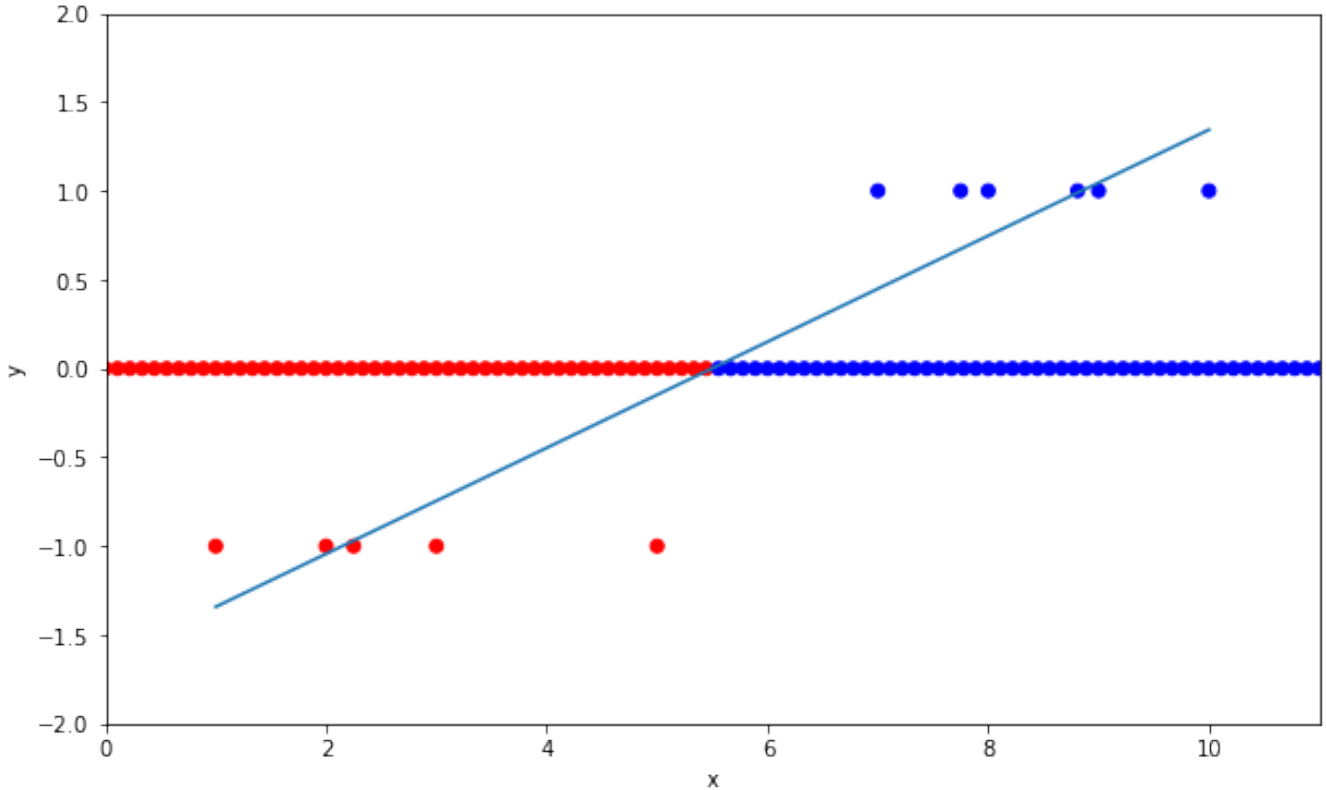
Classificação binária via regressão linear

Podemos determinar $\mathbf{w}^T \mathbf{x}$ via regressão linear

E fazer

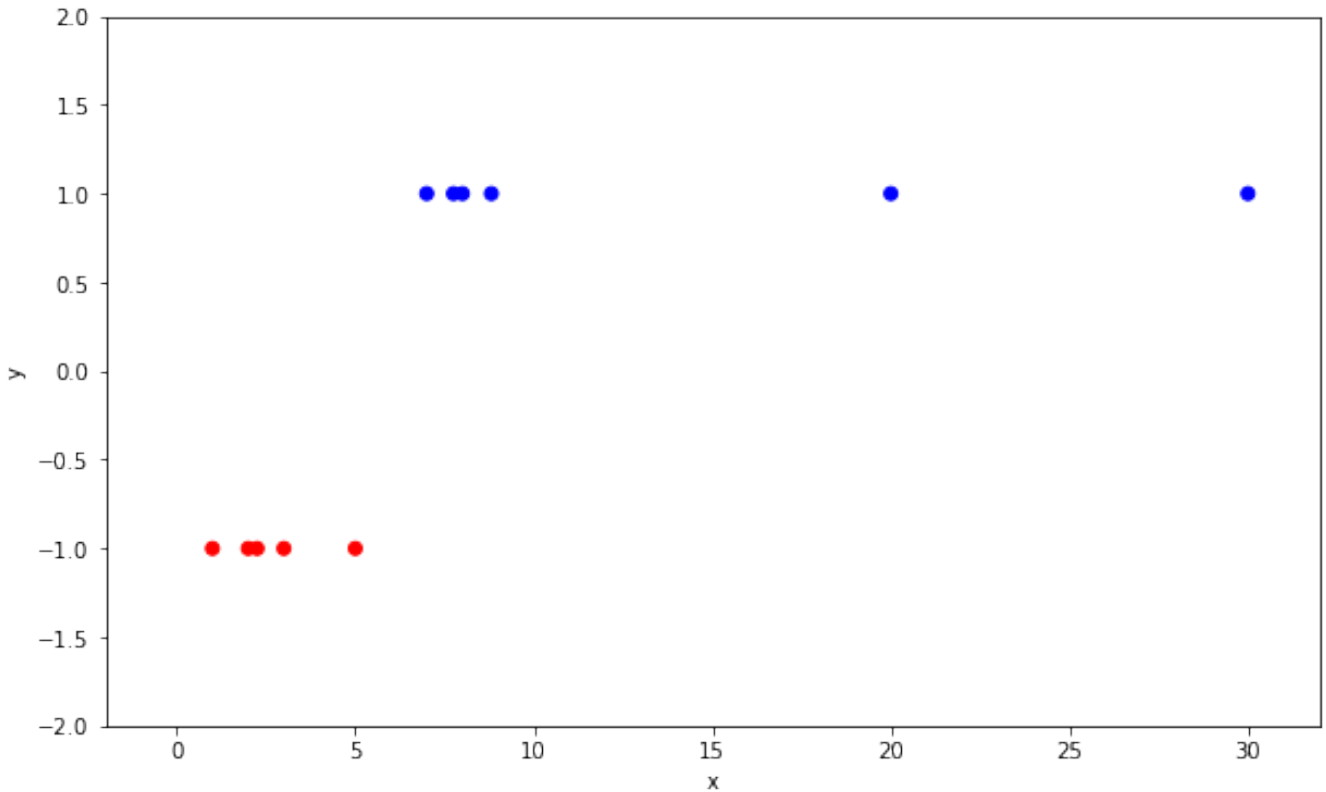
$$\hat{y} = \text{sign}\left(\mathbf{w}^T \mathbf{x}\right)$$

Classificação binária via regressão linear



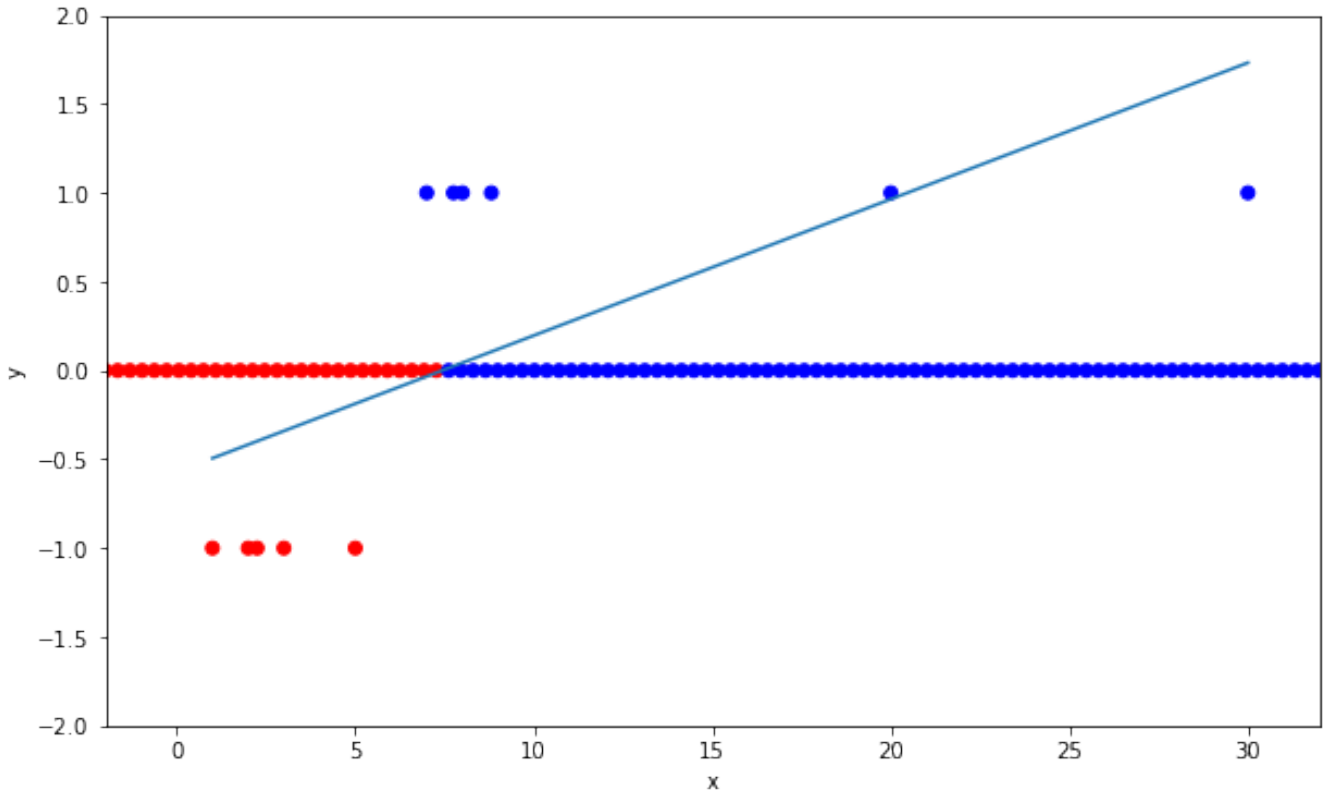
Classificação binária via regressão

Second example: $D_2 = \{1, 2, 2.25, 3, 5, 7, 7.75, 8, 8.81, 20, 30\}$

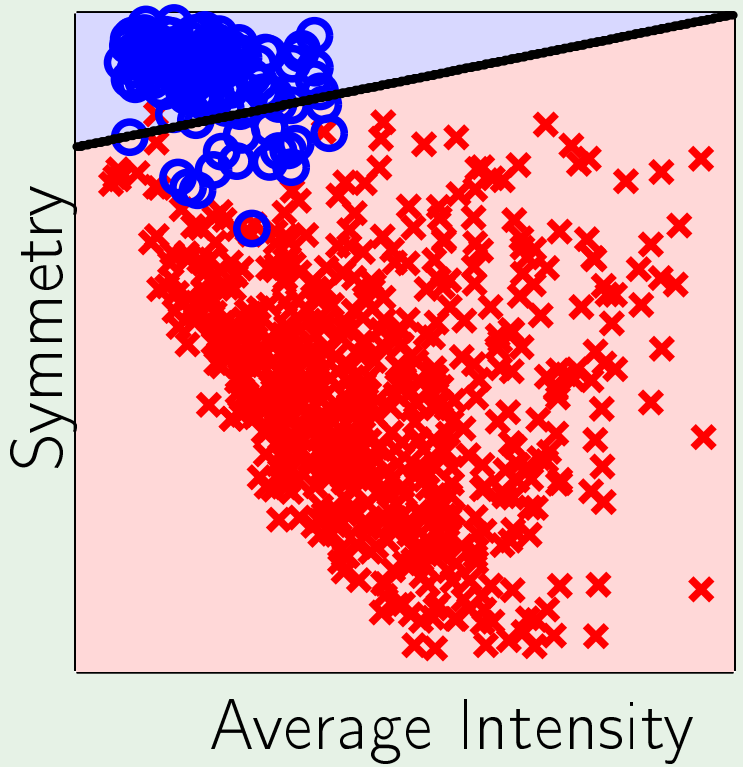


Classificação binária via regressão linear

Os pontos positivos mais a esquerda serão erroneamente classificados como negativos



Linear regression boundary



Discussão

Quando usamos a técnica de regressão em problemas de classificação ...

- Pontos distantes da fronteira de decisão tendem a ter resíduo maior

- Se o espalhamento dos dados em torno da fronteira de decisão ótima não é simétrico, temos um efeito de amostras de uma classe contribuindo mais para a função de erro do que as da outra classe

- Ao se fazer a otimização, ocorre uma espécie de equilíbrio entre as contribuições de ambas as classes, o que resulta em uma fronteira deslocada para "dentro" de uma das classes

Binary classification

Logistic Regression

Classification when we have noisy targets

Na prática, um mesmo exemplo $\mathbf{x}$ pode ora ser observado como sendo da classe $y = 1$ e ora da classe $y = 0$.

É, por exemplo, o caso da altura. Supondo que $\mathbf{x}$ é a altura de uma pessoa, é bem provável que uma mesma altura seja observada mais de uma vez. É também bem provável que, dependendo da altura, ora a pessoa seja do sexo masculino, ora seja do sexo feminino.

Neste sentido, a distribuição que de fato nos interessa é $P(y|\mathbf{x})$

Modelos podem ser desenhados para, dado x, predizer $P(y|\mathbf{x})$, em vez de atribuir x a uma classe $\hat{y}$

A classe final $\hat{y}$ pode então ser definido em função das predições $P(y_j|\mathbf{x})$ (para todas as classes j possíveis) e do contexto de aplicação

Cenário de noisy target

Suponha classificação binária: $y \in \{-1, +1\}$

Observações $(\mathbf{x}, y)$ seguem uma distribuição $P(X, Y)$

O target que queremos aprender não é mais uma função que determina y para cada $\mathbf{x}$

Nosso target agora é uma distribuição $P(y|\mathbf{x})$

Caracterização de $P(y|\mathbf{x})$

Se definirmos $f(\mathbf{x}) = P(y = +1|\mathbf{x})$, a distribuição $P(y|\mathbf{x})$ pode ser caracterizada por:

$$P(y|\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \text{se } y = +1, \\ 1 - f(\mathbf{x}), & \text{se } y = -1 \end{cases}$$

(note que, alternativamente, poderíamos ter definido $f(\mathbf{x}) = P(y = -1|\mathbf{x})$)

Note que não temos acesso a $f(\mathbf{x})$; só sabemos que y segue a distribuição $P(y|\mathbf{x})$

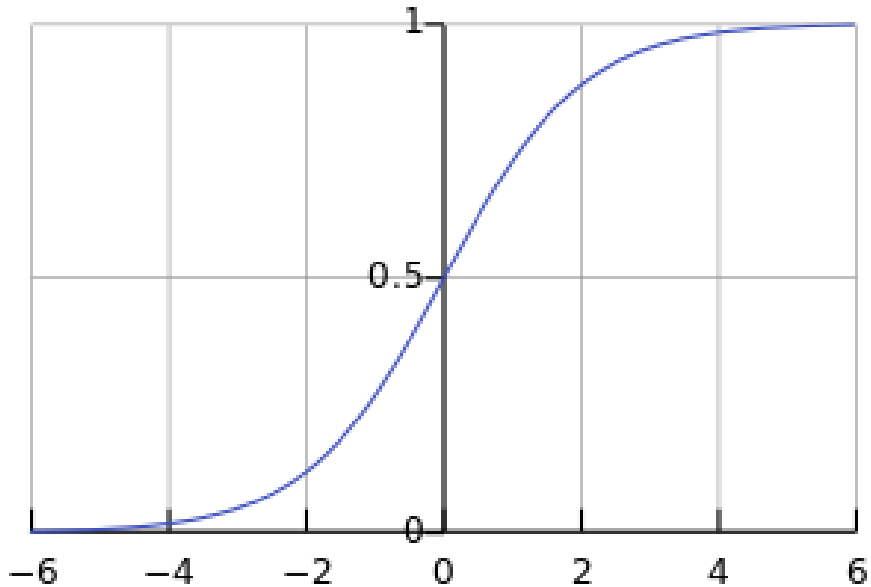
Se conseguirmos aprender $f(\mathbf{x})$, teremos aprendido $P(y|\mathbf{x})$

Função logística ou sigmóide

Ideia é considerarmos hipóteses do tipo

$$h_{\mathbf{w}}(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x})$$

$$\theta(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1}$$



$$0 \leq \theta(z) \leq 1 \quad \Longrightarrow \quad 0 \leq h_{\mathbf{w}}(\mathbf{x}) \leq 1$$

Truque

Supondo que $h_{\mathbf{w}}(\mathbf{x}) \approx f(\mathbf{x})$, é natural supor que

$$\hat{P}(y|\mathbf{x}) \approx \begin{cases} h_{\mathbf{w}}(\mathbf{x}), & \text{se } y = +1, \\ 1 - h_{\mathbf{w}}(\mathbf{x}), & \text{se } y = -1 \end{cases}$$

é uma boa aproximação de $P(y|\mathbf{x})$

Note que $1 - \theta(z) = \theta(-z)$

Usando isso, podemos escrever

$$\hat{P}(y|\mathbf{x}) = \theta(y\,\mathbf{w}^T\mathbf{x})$$

(isto é apenas um truque para não precisar tratar $\hat{P}(y|\mathbf{x})$ dividido em dois casos)

Regressão logística

$h(\mathbf{x}) = \theta(\mathbf{w}^T \mathbf{x})$ pode ser interpretada como probabilidade

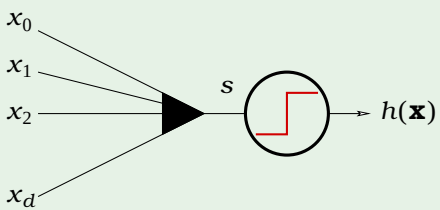
(já que $0 \leq \theta(z) \leq 1$)

A third linear model

$$s = \sum_{i=0}^{d} w_i x_i$$

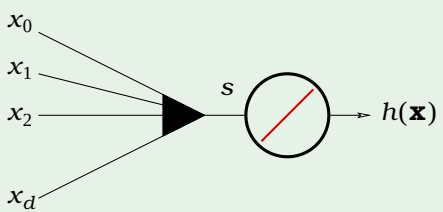
linear classification

$$h(\mathbf{x}) = \text{sign}(s)$$



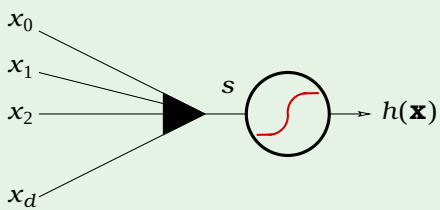
linear regression

$$h(\mathbf{x}) = s$$



logistic regression

$$h(\mathbf{x}) = \theta(s)$$



Aprendizado do target $f(\mathbf{x}) = P(y = 1|\mathbf{x})$

Dados disponíveis:

$$\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)}) \in X \times Y, n = 1, \ldots, N\}$$

Dada uma observação $(\mathbf{x}, y)$, nós estamos supondo que y segue uma distribuição $P(y|\mathbf{x}) = \theta(y\,\mathbf{w}^T\mathbf{x})$

QUEREMOS determinar $\mathbf{w}$ que irá definir a distribuição (desconhecida) $P(y|\mathbf{x}) = \theta(y\,\mathbf{w}^T\mathbf{x})$ que deu origem às observações em $\mathcal{D}$

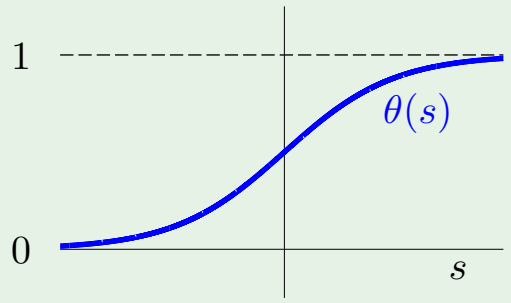
Princípio da máxima verossimilhança

Dentre todas as distribuições $\theta(y\,\mathbf{w}^T\mathbf{x})$, qual é a que gerou D ?
É aquela que maximiza a probabilidade conjunta

$$\prod_{n=1}^{N} P(y^{(n)}|\mathbf{x}^{(n)}) = \prod_{n=1}^{N} \theta(y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)})$$

Formula for likelihood

$$P(y \mid \mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{for } y = +1; \\ 1 - h(\mathbf{x}) & \text{for } y = -1. \end{cases}$$



Substitute $h(\mathbf{x}) = \theta(\mathbf{w}^\mathsf{T}\mathbf{x})$, noting $\theta(-s) = 1 - \theta(s)$

$$P(y \mid \mathbf{x}) = \theta(y\ \mathbf{w}^\mathsf{T}\mathbf{x})$$

Likelihood of $\mathcal{D} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ is

$$\prod_{n=1}^{N} P(y_n \mid \mathbf{x}_n) = \prod_{n=1}^{N} \theta(y_n \mathbf{w}^\mathsf{T}\mathbf{x}_n)$$

Problema de otimização a ser resolvido

Encontrar $\mathbf{w}$ que maximiza

$$\prod_{n=1}^{N} \theta(y^{(n)} \mathbf{w}^T \mathbf{x}^{(n)})$$

Ou, equivalentemente, maximiza

$$\frac{1}{N} \ln \left(\prod_{n=1}^{N} \theta(y^{(n)} \mathbf{w}^T \mathbf{x}^{(n)}) \right)$$

Ou, ainda, minimiza

$$-\frac{1}{N} \ln \left(\prod_{n=1}^{N} \theta(y^{(n)} \mathbf{w}^T \mathbf{x}^{(n)}) \right)$$

Quero w que minimiza

$$-\frac{1}{N} \ln \left(\prod_{n=1}^{N} \theta(y^{(n)} \mathbf{w}^T \mathbf{x}^{(n)}) \right)$$

$$-\frac{1}{N} \sum_{n=1}^{N} \ln \left(\theta(y^{(n)} \mathbf{w}^T \mathbf{x}^{(n)}) \right) \quad (\text{ pois } \ln \textstyle\prod a_i = \sum \ln a_i \ )$$

$$\frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1}{\theta(y^{(n)} \mathbf{w}^T \mathbf{x}^{(n)})} \right) \quad (\text{ pois } \ln \tfrac{1}{a} = -\ln a \ )$$

$$\frac{1}{N} \sum_{n=1}^{N} \ln \left(1 + e^{-y^{(n)} \mathbf{w}^T \mathbf{x}^{(n)}} \right) \quad (\text{ pois } \tfrac{1}{\theta z} = \tfrac{1}{\frac{1}{1+e^{-z}}} \ )$$

Função de custo na regressão logística

$$E_{in} = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ln\left(1 + e^{-y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}}\right)}_{err(y^{(n)}, \hat{y}^{(n)})}$$

Interpretação:

Se $y^{(n)}$ e $\mathbf{w}^T\mathbf{x}^{(n)}$ concordam, o expoente em $e^{-y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}}$ é negativo $\rightsquigarrow$ $err(y^{(n)}, \hat{y}^{(n)})$ tende a ser próximo de zero

Se $y^{(n)}$ e $\mathbf{w}^T\mathbf{x}^{(n)}$ discordam, o expoente em $e^{-y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}}$ é positivo $\rightsquigarrow$ $err(y^{(n)}, \hat{y}^{(n)})$ tende a ser grande

Até aqui vimos a formulação usada pelo Mostafa.

Isto é $y \in \{-1, +1\}$

Porém parece ser mais comum a formulação que considera $y \in \{0, 1\}$ (a seguir)

Cross-entropy error

Em vez de $Y = \{-1, +1\}$ podemos considerar $Y = \{0, 1\}$ e escrever (de novo, isso é um truque para não dividir $P(y|\mathbf{x})$ em dois casos):

$$\begin{aligned} P(y|\mathbf{x}) &= P(y = 1|\mathbf{x})^y \, P(y = 0|\mathbf{x})^{1-y} \\ &= P(y = 1|\mathbf{x})^y \, [1 - P(y = 1|\mathbf{x})]^{1-y} \end{aligned}$$

Função de máxima verossimilhança (tirando o índice (n) para limpar a notação)

$$\begin{aligned} \prod_{(\mathbf{x},y)\in D} P(y|\mathbf{x}) &= \prod_{(\mathbf{x},y)\in D} P(y = 1|\mathbf{x})^y \, [1 - P(y = 1|\mathbf{x})]^{1-y} \\ &\approx \prod_{(\mathbf{x},y)\in D} [\theta(\mathbf{w}^T\mathbf{x})]^y \, [1 - \theta(\mathbf{w}^T\mathbf{x})]^{1-y} \\ &= \prod_{(\mathbf{x},y)\in D} \hat{y}^y \, (1 - \hat{y})^{1-y} \end{aligned}$$

Maximizar

$$\prod_{(\mathbf{x},y)\in D} \hat{y}^y \, (1-\hat{y})^{1-y}$$

é minimizar

$$-\ln \prod_{(\mathbf{x},y)\in D} \hat{y}^y \, (1-\hat{y})^{1-y}$$

Que é equivalente a minimizar

$$-\sum_{(\mathbf{x},y)\in D} \ln\left(\hat{y}^y \, (1-\hat{y})^{1-y}\right)$$

$$-\sum_{(\mathbf{x},y)\in D} \ln(\hat{y}^y) + \ln((1-\hat{y})^{1-y})$$

$$-\sum_{(\mathbf{x},y)\in D} y \ln \hat{y} + (1-y)\ln(1-\hat{y})$$

Cross-entropy loss

$$J(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} y^{(n)} \ln \hat{y}^{(n)} + (1 - y^{(n)}) \ln(1 - \hat{y}^{(n)})$$

Dadas duas distribuições discretas p e q sobre A, cross-entropy é:

$$H(p, q) = -\sum_{a \in A} p(a) \log q(a)$$

Otimização usando gradient descent

Gradient – cost function used by Mostafa

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + e^{-y^{(n)}\,\mathbf{w}^T\mathbf{x}^{(n)}}\right)$$

Vamos analisar $\frac{\partial}{\partial \mathbf{w}}[\ln\left(1 + e^{-y\,\mathbf{w}^T\mathbf{x}}\right)]$

Denotando $\mathbf{s} = -y\mathbf{x}$, queremos calcular $\frac{\partial}{\partial \mathbf{w}}[\ln\left(1 + e^{\mathbf{w}^T\mathbf{s}}\right)]$

Lembrando que $\frac{\partial}{\partial \mathbf{w}}[\ln[f(x)] = \frac{f'(x)}{f(x)}$, temos

$\frac{\partial}{\partial \mathbf{w}}[\ln(1 + e^{\mathbf{w}^T\mathbf{s}})] = \frac{(1+e^{\mathbf{w}^T\mathbf{s}})'}{1+e^{\mathbf{w}^T\mathbf{s}}} = \frac{\mathbf{s}\,e^{\mathbf{w}^T\mathbf{s}}}{1+e^{\mathbf{w}^T\mathbf{s}}} = \mathbf{s}\frac{e^{\mathbf{w}^T\mathbf{s}}}{1+e^{\mathbf{w}^T\mathbf{s}}} = \mathbf{s}\frac{1}{1+e^{-\mathbf{w}^T\mathbf{s}}}$

Logo

$$\frac{\partial}{\partial \mathbf{w}}[\ln\left(1 + e^{-y\,\mathbf{w}^T\mathbf{x}}\right)] = -\frac{y\mathbf{x}}{1 + e^{y\mathbf{w}^T\mathbf{x}}}$$

Gradient – cross-entropy loss case

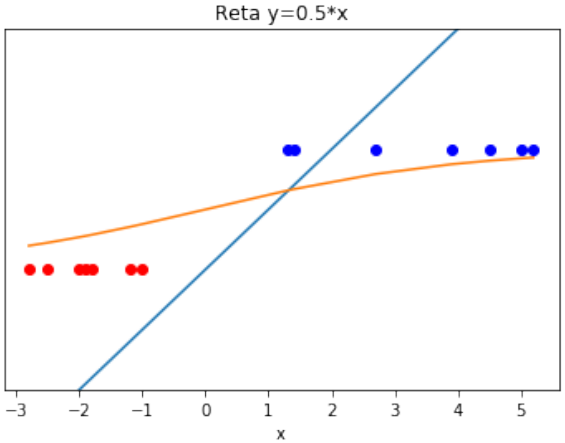
$$J(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} y^{(n)} \ln \hat{y}^{(n)} + (1 - y^{(n)}) \ln(1 - \hat{y}^{(n)})$$

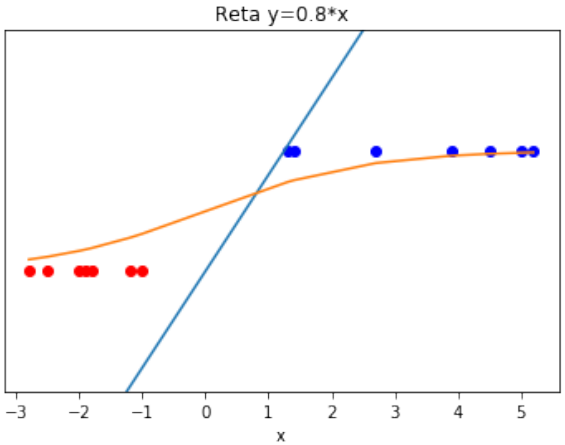
$$\hat{y} = h_{\mathbf{w}}(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$

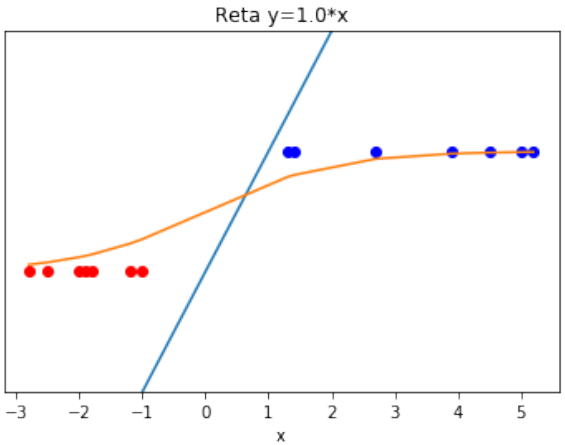
... após alguns cálculos ...

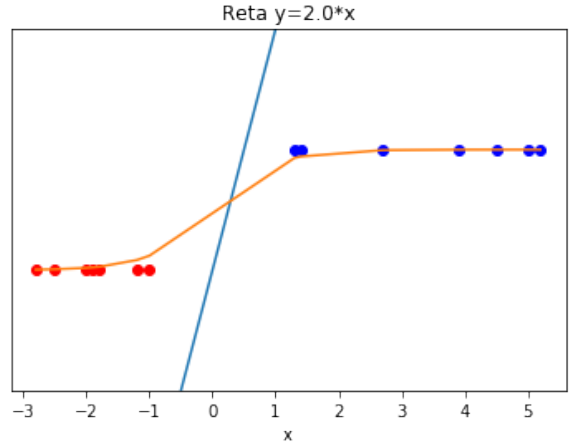
Derivadas parcias: $$\frac{\partial}{\partial w_j} J(\mathbf{w}) = \sum_{n=1}^{N} (\hat{y}^{(n)} - y^{(n)}) x_j^{(n)}$$

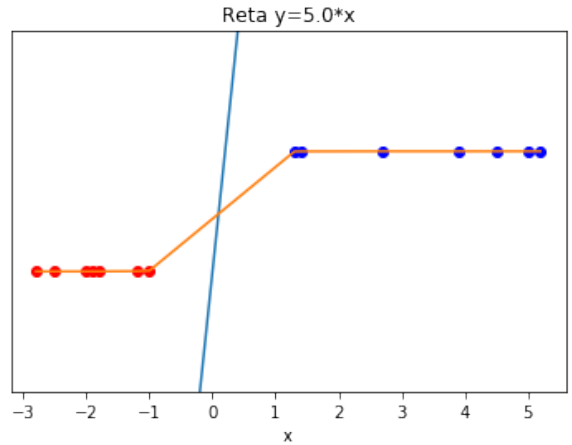
Update de peso: $$\Delta w_j(r) = \sum_{n=1}^{N} (y^{(n)} - \hat{y}^{(n)}) \mathbf{x}_j^{(i)}$$











Reta azul $\mathbf{w}^T \mathbf{x} = w_0 + w_1 x = 0$

Curva laranja: $h(\mathbf{x}) = \theta(\mathbf{w}^T \mathbf{x})$

Contraste entre regressão linear e a logística