Deep Learning

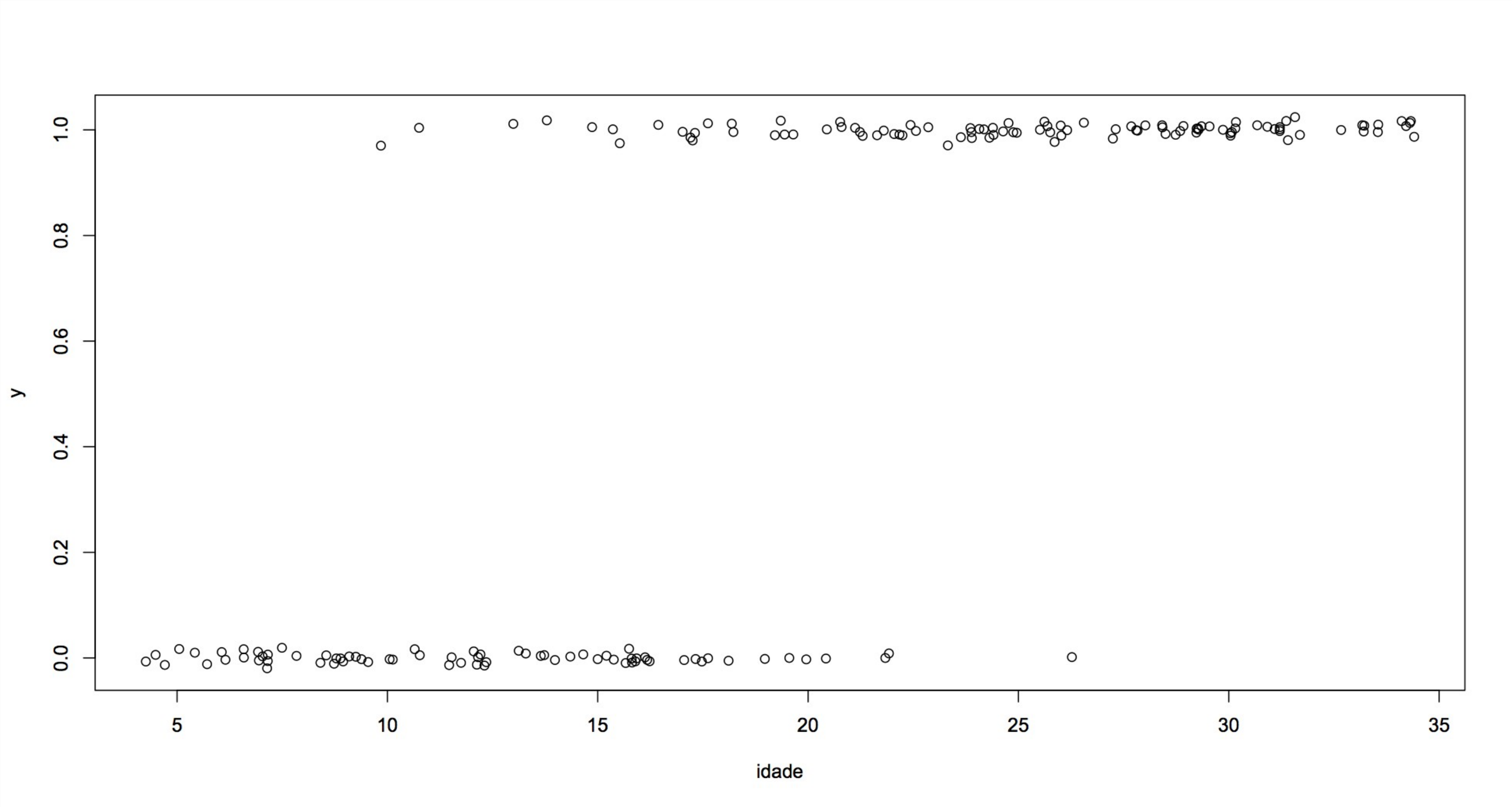
Aula 02 - Semana 01

Renato Assunção - DCC - UFMG

Como escolher a melhor curva logística para ajustar aos dados?

- Várias perguntas:
  - Como obter os coeficientes de uma curva (regressão) logística?
  - Como escolher a "melhor" curva logística? "Melhor" em que sentido?
  - Como avaliar se o modelo logístico é um bom classificador?
  - Como generalizar o modelo se tivermos várias features?
    - E se a probabilidade também da escolaridade da mãe, do sexo da criança, ...
- Roteiro da estrada à frente (aula 02):
  - Método de máxima verossimilhança (ML) e função de custo
  - Otimização: Newton e gradiente descendente
  - Heurística e otimalidade da ML
  - Stochastic gradient descent

173 crianças



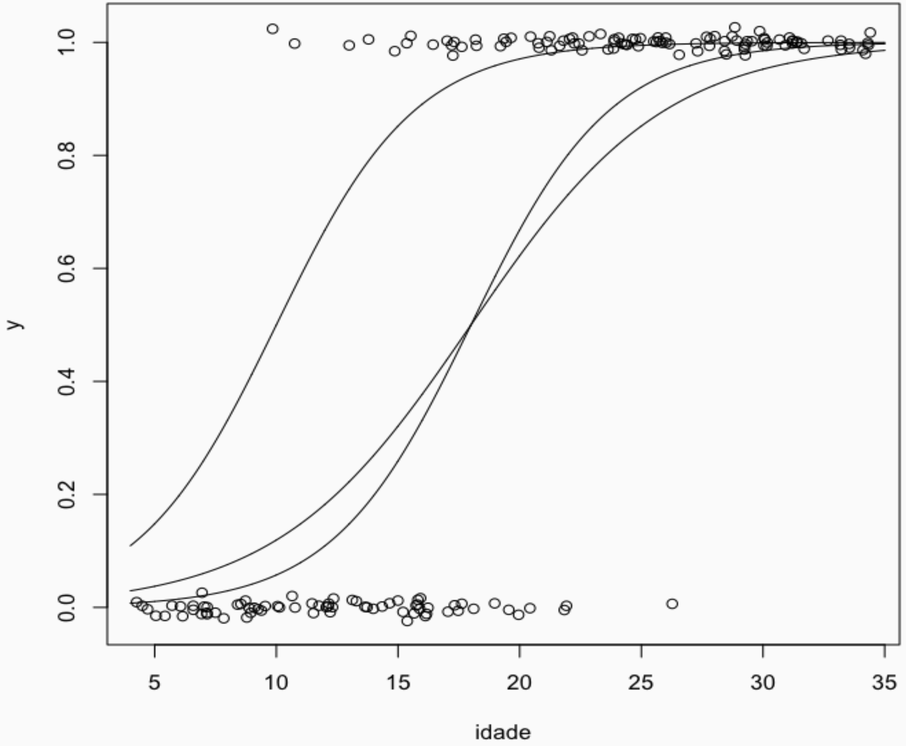
Função logística

- A probabilidade de uma criança com idade x realizar a tarefa é

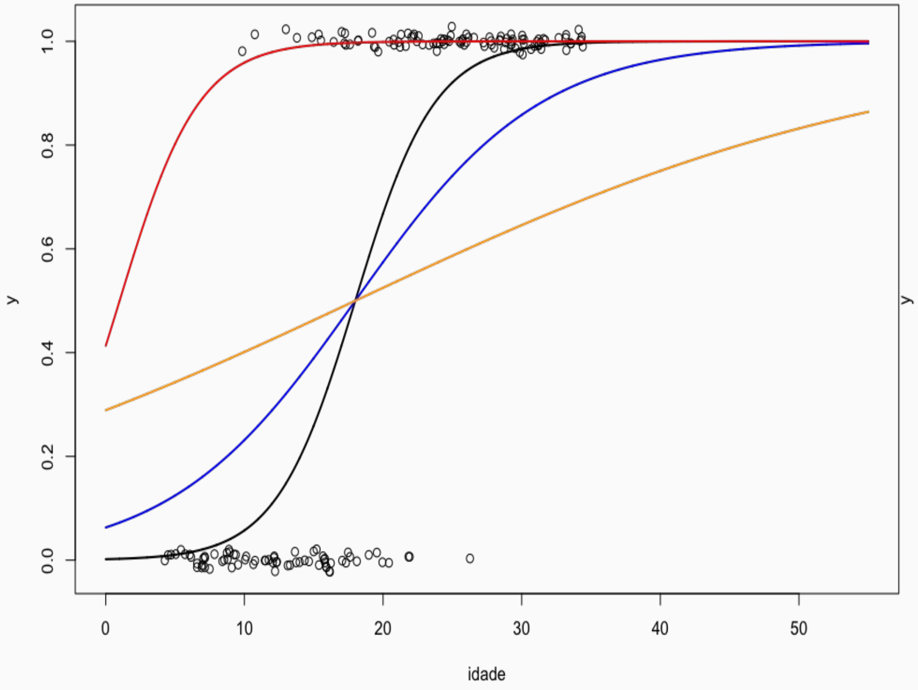
$$\sigma(x) = \frac{1}{1+e^{-(w_0+w_1 x)}}$$

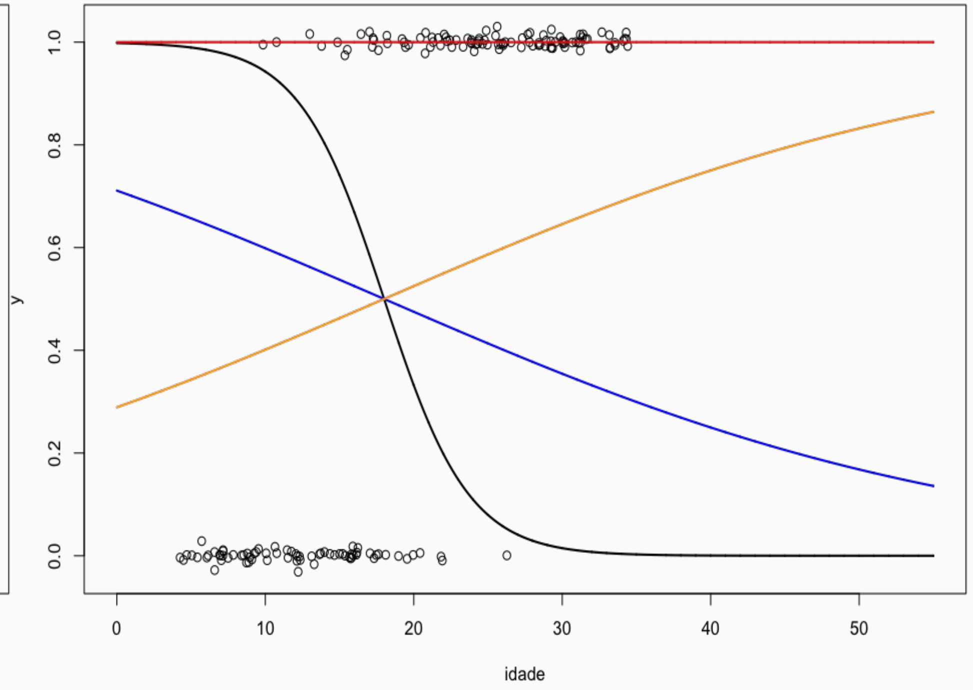
- Como escolher w_0 e w_1 compatíveis com os dados?

- Ideia: escolha w_0 e w_1 de tal forma que os dados realmente observados possam ser gerados pelo modelo.



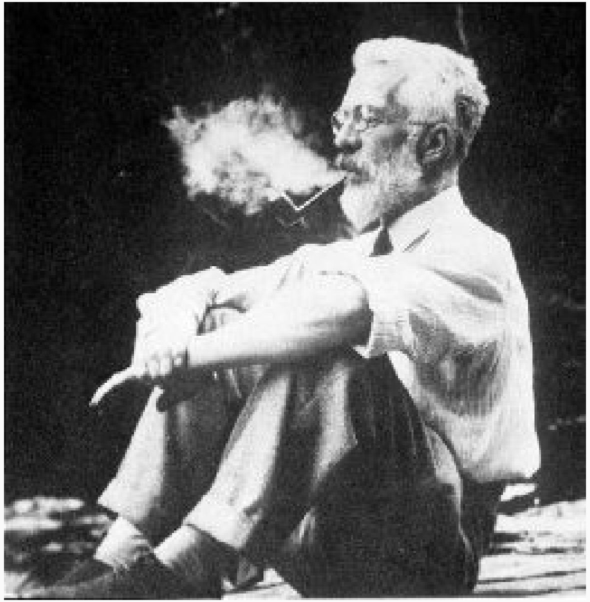
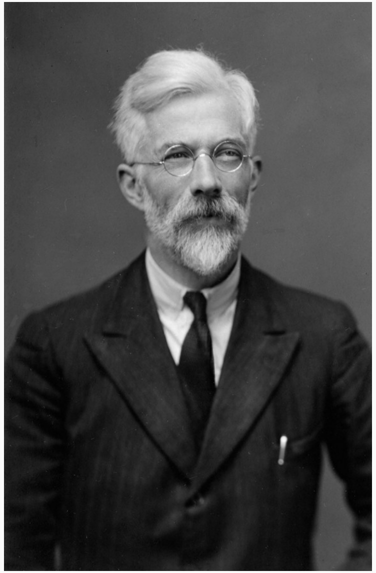
Diferentes parâmetros, diferentes curvas.





Ideia: Algumas das curvas são "compatíveis" com os dados.
Algumas curvas são <u>verossímeis</u> como modelo gerador dos dados observados.

- Método de máxima verossimilhança → para estimar parâmetros ou coeficientes com dados estatísticos
- Foi criado por Sir Ronald Fisher (1890 - 1962), o maior estatístico que já existiu.

E a luz se fez em 1922

- Fisher foi uma espécie de Isaac Newton da estatística, responsável pelos principais conceitos e resultados da inferência estatística, usados até hoje.
- Suas ideias principais foram publicadas de uma só vez, num artigo de 1922, *On the mathematical foundations of theoretical statistics.*
- Alguns dos principais conceitos e resultados usados até hoje:
  - verossimilhança,
  - suficiência
  - vício e eficiência de estimação
- são desse artigo maravilhoso (ele tinha 32 anos de idade).

IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by Dr. E. J. Russell, F.R.S.

Received June 25,—Read November 17, 1921.

Contents.

Definitions.

Centre of Location.—That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (9.)

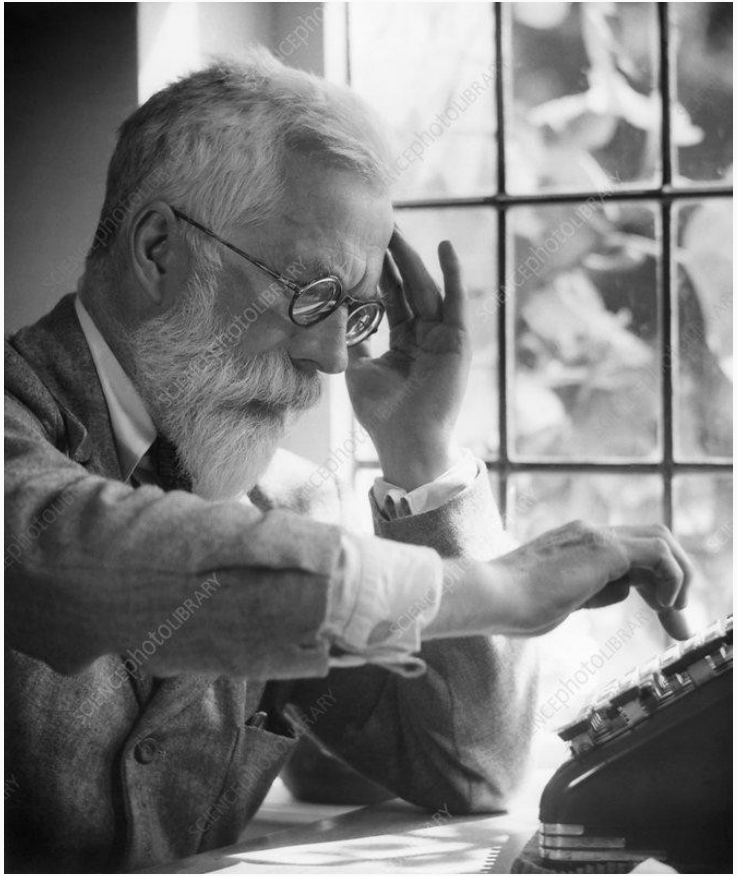
Consistency.—A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

Distribution.—Problems of distribution are those in which it is required to calculate the distribution of one, or the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

Efficiency.—The efficiency of a statistic is the ratio (usually expressed as a percentage) which its intrinsic accuracy bears to that of the most efficient statistic possible. It

Mais um pouco de Fisher

- Fisher foi também um maiores geneticistas que já existiu
    - junto com Sewall Wright e Haldane, é responsável por juntar de forma coerente a teoria da evolução de Darwin e a teoria genética de Mendel (um quebra-cabeça complicado em 1920)
- Criador de:
    - teoria e prática do planejamento de experimentos (aleatorização, blocagem, quadrados-latinos, etc)
    - Análise de regressão linear (p-valores)
    - PCA
    - Análise discriminante
    - Teoria de valores extremos, etc etc etc etc etc



Verossimilhança = Likelihood

- Vimos algumas curvas logísticas "extremas".
- Dificilmente elas poderiam ter gerado os dados das crianças.
- Fisher: estas curvas extremas não são ***verossímeis***.
    - vero: verdadeiro, real, autêntico;
    - símil: semelhante, similar.
- algo é verossímil se parece verdadeiro,
    - se não repugna à verdade,
    - se é semelhante à verdade,
    - se é coerente o suficiente para se passar por verdade.
- Ao dizer que algo é verossímil, ***não*** dizemos que é verdadeiro.
- Verossímil = *parece verdadeiro* pois está de acordo com todas as evidências disponíveis

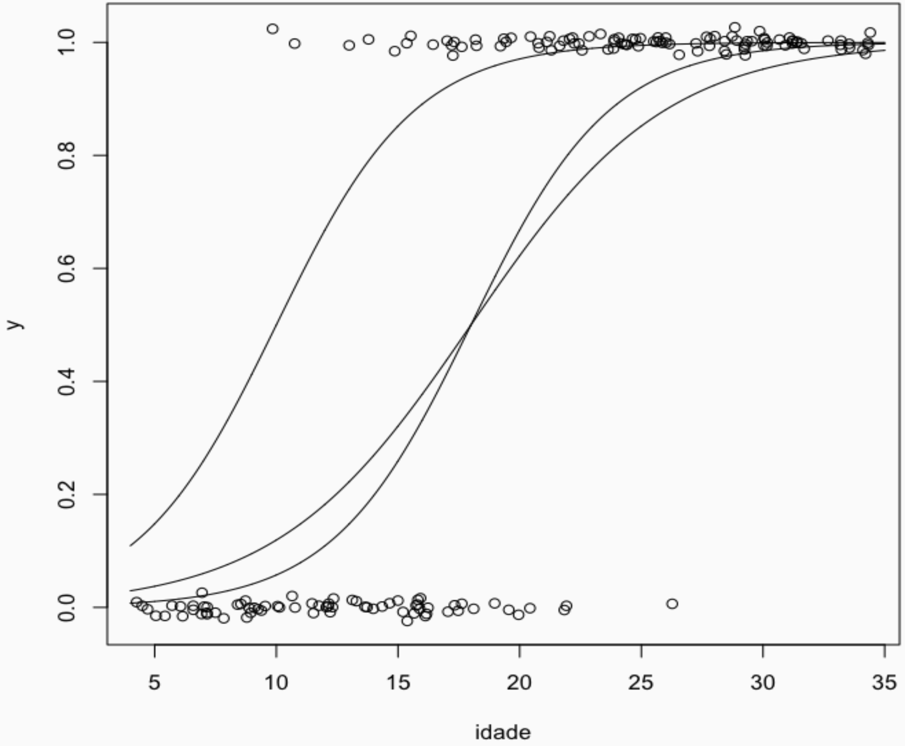
A verossimilhança do modelo logístico

- A probabilidade de uma criança com idade x realizar a tarefa é

$$\sigma(x) = \frac{1}{1+e^{-(w_0+w_1 x)}}$$

- Vamos fixar w_0 e w_1 → fixar uma curva

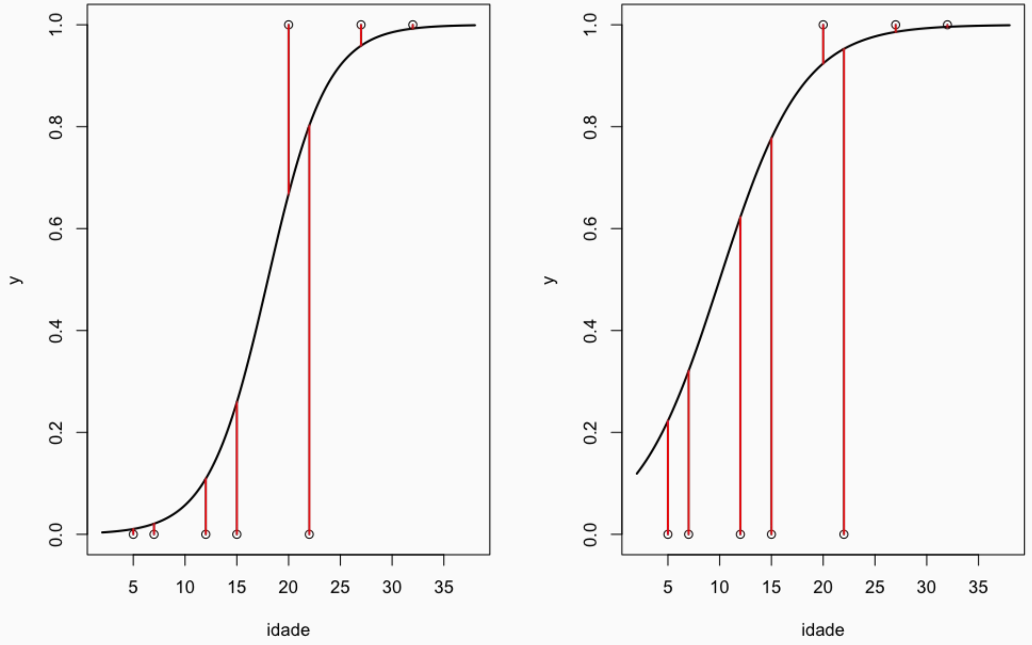
- Para esta curva fixada, obtenha a probabilidade de gerar os sucessos e fracassos realmente observados.



Duas curvas e suas probabilidades

- Para cada curva possível:
  - calcular a probabilidade de gerar os valores 0 ou 1 *realmente observados*
  - Multiplicar estas probabilidades (regra de indep de eventos: as crianças agem independentemente)
  - Obter a probabilidade para cada curva
  - Para qual curva esta probabilidade é máxima?

- Fazer exemplo no quadro comparando duas curvas com 5 pontos.



A função de verossimilhança

- Temos 5 crianças com idades x iguais a 5, 12, 22, 25, 30
- Os y's correspondentes são 0, 1, 0, 1, 1
- Se w_0 = -6.3 e w_1 = 0.35, obtenha a probabilidade de gerar os y's acima com o modelo logístico
- Para cada criança e para estas escolhas de w_0 e w_1, esta probabilidade é

$$\sigma(x) = \frac{1}{1+e^{-(-6.3+0.35x)}}$$

- Vamos refazer este cálculo obtendo esta probabilidade com diferentes valores de w_0 e w_1
- Esta probabilidade será uma função de w_0 e w_1

$$L(w_0, w_1) = \mathbb{P}(Y_1 = 0, Y_2 = 1, Y_3 = 0, Y_4 = 1, Y_5 = 1 | w_0, w_1)$$

Obtendo a verossimilhança para $w_0 = -6.3 \quad \text{e} \quad w_1 = 0.35$

- Sejam $w_0 = -6.3 \quad \text{e} \quad w_1 = 0.35$
- Vamos obter

$$L(-6.3, 0.35) = \mathbb{P}(Y_1 = 0, Y_2 = 1, Y_3 = 0, Y_4 = 1, Y_5 = 1 | w_0 = -6.3, w_1 = 0.35)$$

- O resultado de uma criança (sucesso ou fracasso) não afeta o resultados das demais crianças. São eventos independentes.

$$L(-6.3, 0.35) = \mathbb{P}(Y_1 = 0 | w_0 = -6.3, w_1 = 0.35)\mathbb{P}(Y_2 = 1 | w_0 = -6.3, w_1 = 0.35) \times$$
$$\times \mathbb{P}(Y_3 = 0 | w_0 = -6.3, w_1 = 0.35)\mathbb{P}(Y_4 = 1 | w_0 = -6.3, w_1 = 0.35)\mathbb{P}(Y_5 = 1 | w_0 = -6.3, w_1 = 0.35)$$

- Precisamos calcular cada uma das 5 probabilidades na expressão acima

Obtendo a verossimilhança para $w_0 = -6.3$ e $w_1 = 0.35$

- Queremos $L(-6.3, 0.35) = \mathbb{P}(Y_1 = 0, Y_2 = 1, Y_3 = 0, Y_4 = 1, Y_5 = 1 | w_0 = -6.3, w_1 = 0.35)$
- Isto é igual a

$$L(-6.3, 0.35) = \mathbb{P}(Y_1 = 0 | w_0 = -6.3, w_1 = 0.35)\mathbb{P}(Y_2 = 1 | w_0 = -6.3, w_1 = 0.35) \times$$
$$\times \mathbb{P}(Y_3 = 0 | w_0 = -6.3, w_1 = 0.35)\mathbb{P}(Y_4 = 1 | w_0 = -6.3, w_1 = 0.35)\mathbb{P}(Y_5 = 1 | w_0 = -6.3, w_1 = 0.35)$$

- Temos $\mathbb{P}(Y = 1 | w_0 = -6.3, w_1 = 0.35) = \frac{1}{1+e^{-(-6.3+0.35x)}}$

- e $\mathbb{P}(Y = 0 | w_0 = -6.3, w_1 = 0.35) = 1 - \frac{1}{1+e^{-(-6.3+0.35x)}} = \frac{1}{1+e^{-6.3+0.35x}}$

- A diferença entre as duas probabilidades acima está no expoente da exponencial

Obtendo a verossimilhança para $w_0 = -6.3 \quad \text{e} \quad w_1 = 0.35$

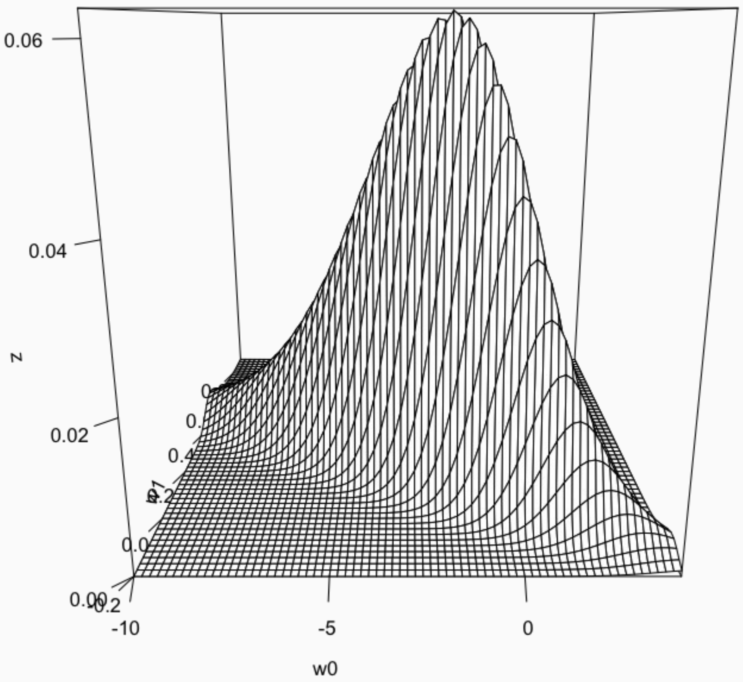
- Temos 5 crianças com idades x iguais a 5, 12, 22, 25, 30
- A verossimilhança para $w_0 = -6.3 \quad \text{e} \quad w_1 = 0.35$ é portanto igual ao produto

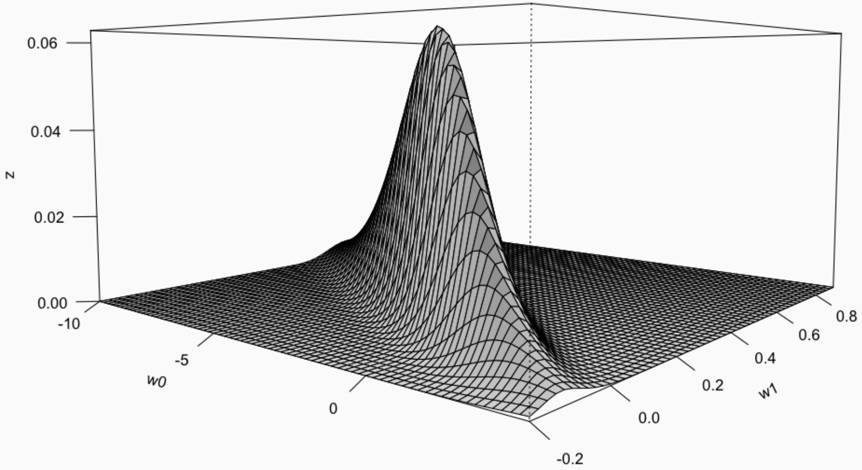
$$
\begin{aligned}
L(-6.3, 0.35) &= (1 - \sigma(5))\, \sigma(12)\, (1 - \sigma(22))\, \sigma(25)\, \sigma(30) \\
&= \frac{1}{1+e^{(-6.3+0.35*5)}} \frac{1}{1+e^{-(-6.3+0.35*12)}} \frac{1}{1+e^{(-6.3+0.35*22)}} \frac{1}{1+e^{-(-6.3+0.35*25)}} \frac{1}{1+e^{-(-6.3+0.35*30)}} \\
&= 0.01936855
\end{aligned}
$$

- Escrevendo esta expressão como função genérica dos coeficientes w_0 e w_1 temos a função de verossimilhança

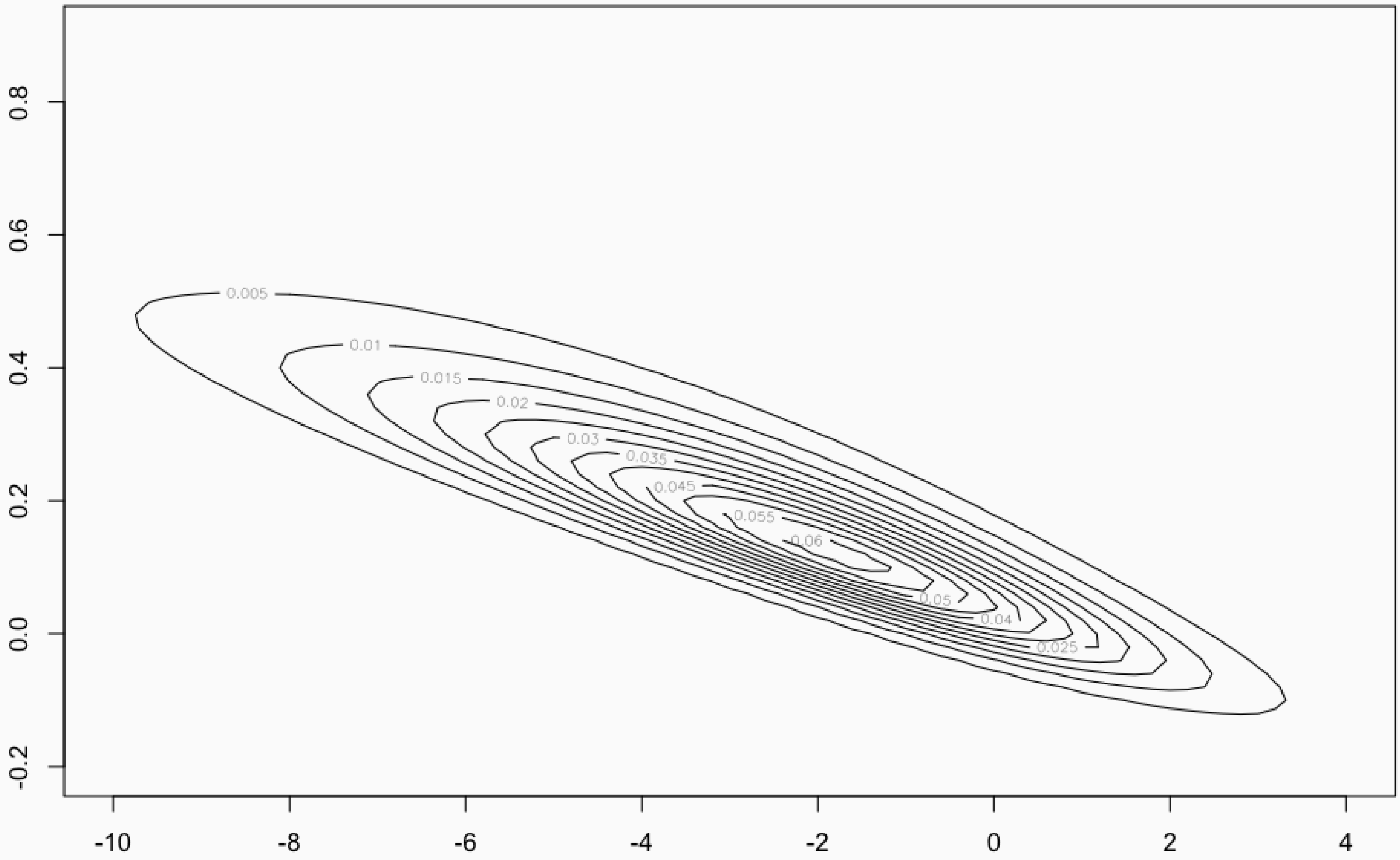
$$
\begin{aligned}
L(w_0, w_1) &= (1 - \sigma(5))\, \sigma(12)\, (1 - \sigma(22))\, \sigma(25)\, \sigma(30) \\
&= \frac{1}{1+e^{(w_0+w_1*5)}} \frac{1}{1+e^{-(w_0+w_1*12)}} \frac{1}{1+e^{(w_0+w_1*22)}} \frac{1}{1+e^{-(w_0+w_1*25)}} \frac{1}{1+e^{-(w_0+w_1*30)}}
\end{aligned}
$$

Função de verossimilhança $L(w_0, w_1)$





Curvas de nível da função de verossimilhança $L(w_0, w_1)$



MLE = Maximum Likelihood Estimator

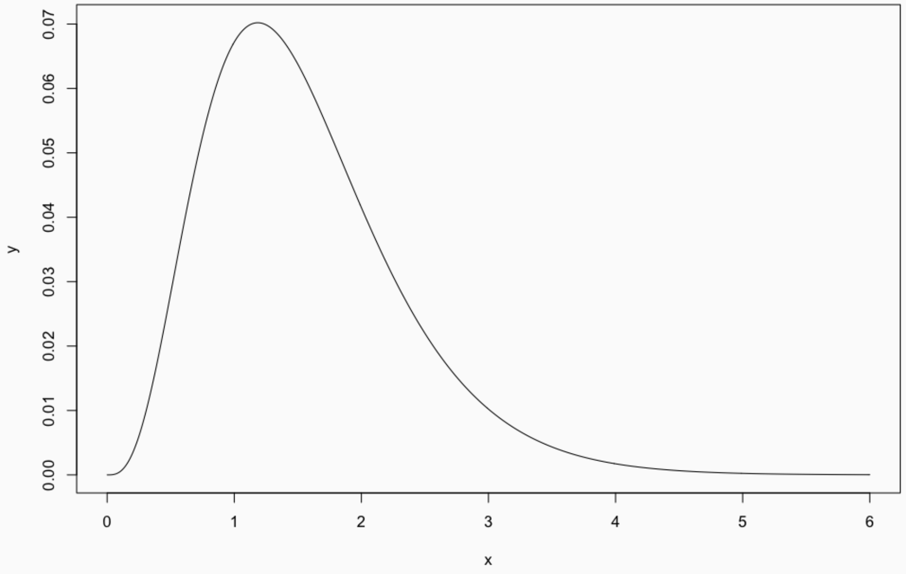
- O MLE é o valor dos coeficientes (w_0, w_1) que maximiza a função de verossimilhança $L(w_0, w_1)$
- Notação: $$(\hat{w}_0, \hat{w}_1) = \underset{(w_0, w_1)}{\arg\max}\, L(w_0, w_1)$$
- Assim, $(\hat{w}_0, \hat{w}_1)$ é o valor dos coeficientes que torna máxima a probabilidade de observar a sequência de dados que realmente observamos
- Pelas curvas de nível do exemplo, vemos que $(\hat{w}_0, \hat{w}_1) \approx (-2, 0.1)$

Obtendo o MLE

- Precisamos de um algoritmo numérico para maximizar $L(w_0, w_1)$
- Método eficiente: método de Newton (ou Newton-Raphson)
- Como funciona?
- Caso uni-dimensional primeiro
- Queremos encontrar o ponto x* tal que f(x*) é o máximo da função f(x)
- Dizemos que x* é o ponto de máximo da função f(x): x* = arg max f(x)
- Como encontrar x*?
    - Derive f(x) obtendo f'(x)
    - Iguale a zero e "resolva" para x → f'(x) = 0 (encontrar a RAIZ desta equação)

Exemplo

- Queremos encontrar o ponto de máximo de $f(x) = x^{3.2}e^{-2.7x}$ para x > 0



Exemplo

- Obtemos a derivada f'(x)

$$f'(x) = 3.2\, x^{2.2} e^{-2.7x} - 2.7\, x^{3.2} e^{-2.7x}$$

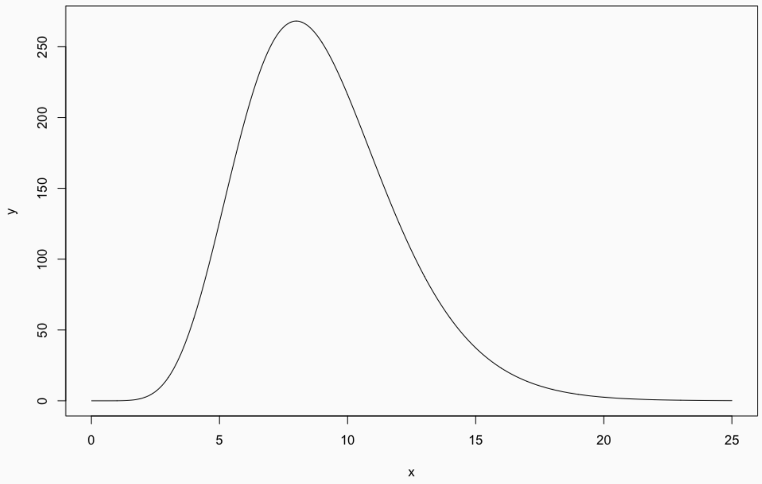
- Iguale f'(x) = 0 e tente "isolar" x. Neste caso, é fácil:

$$\begin{aligned} 3.2\, x^{2.2} e^{-2.7x} &= 2.7\, x^{3.2} e^{-2.7x} \\ 3.2\, x^{2.2} &= 2.7\, x^{3.2} \\ \frac{3.2}{2.7} &= \frac{x^{3.2}}{x^{2.7}} \\ 1.185185 &= x^{0.5} \\ 1.404664 &= x \end{aligned}$$

Exemplo

- Na maioria das vezes não conseguiremos isolar x:

$$f(x) = \frac{x^8}{21(e^x - 1)^{11}}$$



- com derivada

$$f'(x) = \frac{8\,x^7}{21(e^x - 1)^{11}} - \frac{231\,x^8 (e^x - 1) e^x}{441\,(e^x - 1)^{22}}$$

- Não tem "isolar" x para obter o ponto de máximo

Um primeiro passo: tomar log(f(x))

- Likelihood = probabilidade de vários dados
- Usualmente (quase sempre) ela será um PRODUTO de várias funções
- Considere o que é mais fácil derivar:
  - $f(x) = h(x) * g(x) * k(x)$
  - $f(x) = h(x) + g(x) + k(x)$
- Derivada de produtos será uma longa expressão:
  - $f'(x) = h'(x) * g(x) * k(x) + h(x) * g'(x) * k(x) + h(x) * g(x) * k'(x)$
  - $f'(x) = h'(x) + g'(x) + k'(x)$

Primeiro passo: tomar log

- $Log(h(x) * g(x) * k(x)) = Log(h(x)) + Log(g(x)) + Log(k(x))$ ← derivada + simples
- Mas faz sentido?? Queremos max $L(w_0, w_1)$ mas obtemos max $\log(L(w_0, w_1))$
- Na verdade, <u>não queremos</u> max $L(w_0, w_1)$
- Queremos … arg max $L(w_0, w_1)$
- E …. arg max $L(w_0, w_1)$ = arg max $\log(L(w_0, w_1))$
- Por quê?
    - Porque log é função monótona: se $x < y$ então $\log(x) < \log(y)$
    - Assim, se $f(x) < f(x^*)$ para todo $x \neq x^*$ então $\log(f(x)) < \log(f(x^*))$
        - Se x^* maximiza $f(x)$ então x^* também maximiza $\log(f(x))$

Em suma, tome logs

$$(\hat{w}_0, \hat{w}_1) = \arg\max_{(w_0, w_1)} L(w_0, w_1) = \arg\max_{(w_0, w_1)} \log(L(w_0, w_1))$$

- Em conclusão,
- Uma vantagem adicional: estabilidade numérica.
  - probabilidades estão 0 e 1.
  - Multiplicar muitas probabilidades → underflow (< precisão da máquina)
  - Tomar logs diminui substancialmente este problema.
- Em suma, vamos calcular o MLE buscando o máximo do LOG da função de verossimilhança
- Como fazer isto numericamente?

Achar o máximo de g(x) = achar raiz de f(x)

- Achar o máximo de g(x) → pontos onde g'(x) = 0
- Chame g'(x) = f(x)
- Queremos achar as raízes da equação f(x) = 0
- Explicação intuitiva: como Newton deve ter pensado??
- Animação: https://en.wikipedia.org/wiki/Newton%27s_method
- Valor inicial: x_0
- Iterar até convergir: $$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$
- Regras de parada: $|x_{n+1} - x_n| < \varepsilon$ ou $\frac{|x_{n+1} - x_n|}{|x_n|} < \varepsilon$

Achar o máximo de g(x) = achar raiz de f(x)

- Como g'(x) = f(x) ...
- a regra de iteração

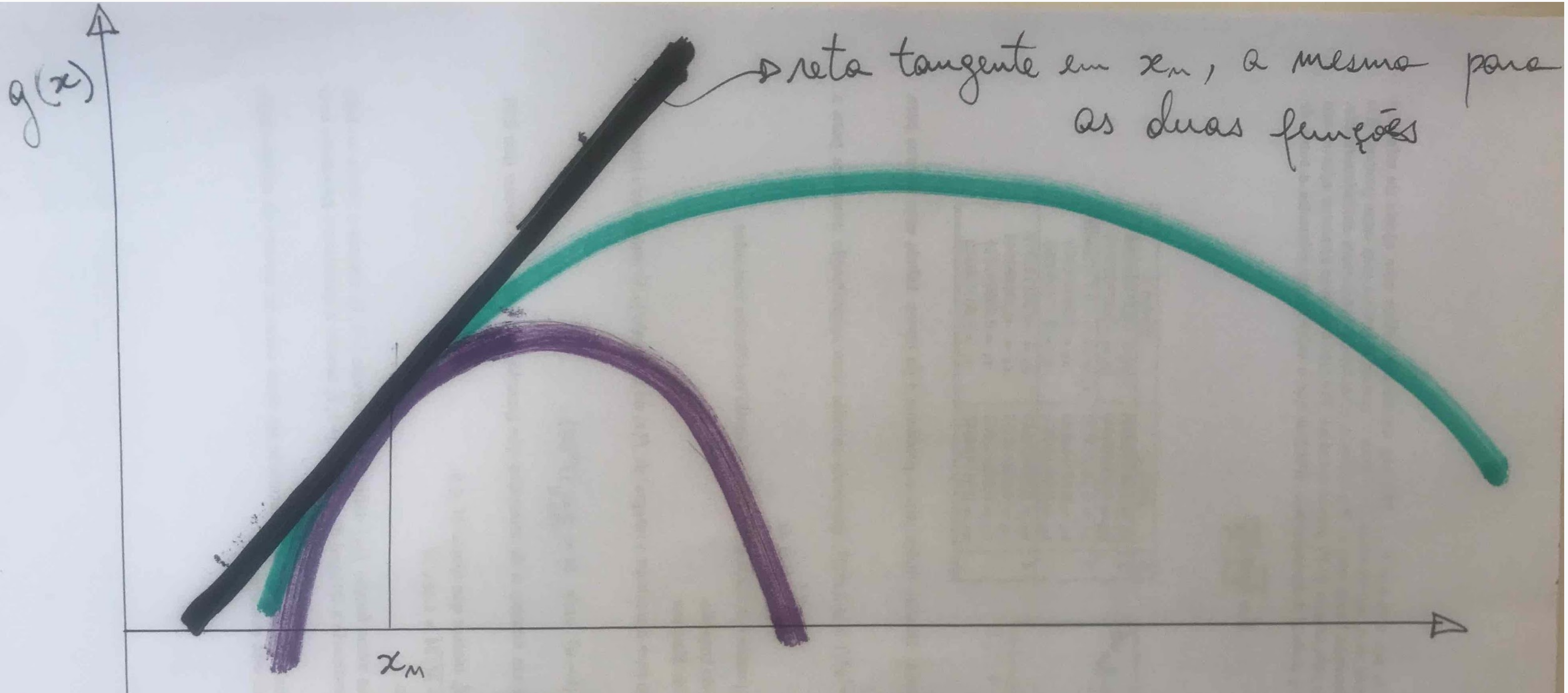
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

- significa

$$x_{n+1} = x_n - \frac{g'(x_n)}{g''(x_n)}$$

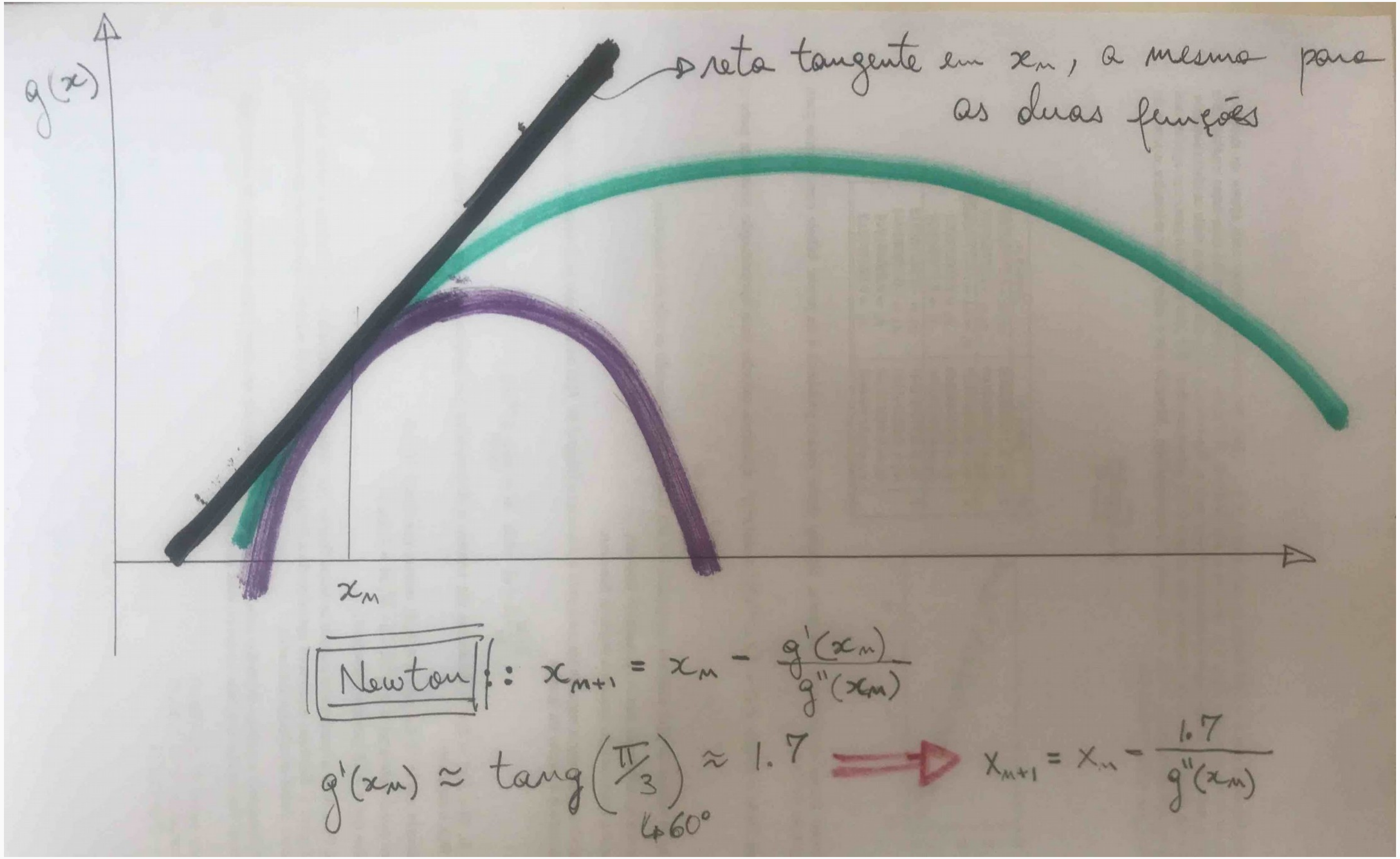
- Vamos ver intuitivamente, o papel de cada termo na fórmula acima:
  - estando em x_n, para que lado andar? para a direita ou para a esquerda?
  - Decidindo para que lado andar, quanto devemos andar?
    - resposta depende de g'
    - e depende também de g''

Explicação intuitiva do método de Newton

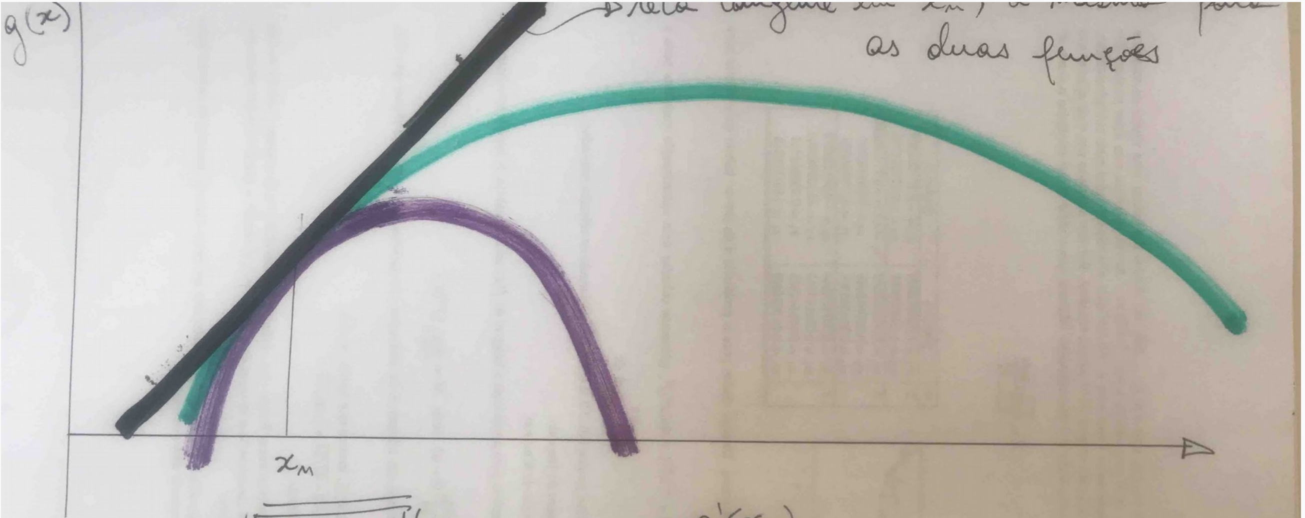


$$\boxed{\text{Newton}}: \; x_{n+1} = x_n - \frac{g'(x_n)}{g''(x_n)}$$

Explicação intuitiva do método de Newton



Explicação intuitiva do método de Newton



$\boxed{\text{Newton}}$: $x_{n+1} = x_n - \dfrac{g'(x_n)}{g''(x_n)}$

$g'(x_n) \approx \tan g\left(\frac{\pi}{3}\right) \approx 1.7 \Longrightarrow x_{n+1} = x_n - \dfrac{1.7}{g''(x_n)}$

↳ 60°

$g''(x_n) = ?$

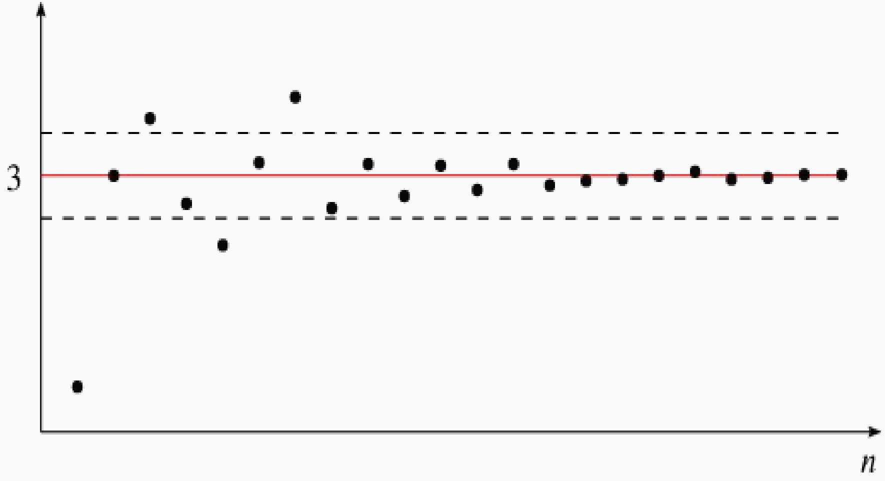
CURVA POSSUI $g''(x_n) \approx 0 \rightarrow x_{n+1} = x_n - \dfrac{1.7}{-0.2} = x_n + 8.5$

CURVA POSSUI $g''(x_n) \ll 0 \rightarrow x_{n+1} = x_n - \dfrac{1.7}{-8.6} = x_n + 0.19$

Convergência de método de Newton

- Grosseiramente, quando converge, o faz rapidamente
- Dobra o no. de casas decimais corretas a cada iteração

- Mas … nem sempre converge
- Existem algumas condições que garantem convergência mas elas em geral não são válida em DL



Generalizando para n features

- Queremos achar o máximo de uma função com mais de uma variável.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \quad \text{um vetor-coluna } n \times 1$$

- Temos uma função

$$\begin{aligned} g : \mathbb{R}^n &\mapsto \mathbb{R} \\ \mathbf{x} &\to g(\mathbf{x}) \end{aligned}$$

Exemplos

$$g(w_0, w_1, w_2) = (w_0^2 + w_1^2 + w_2^2 - 2w_1 w_2)e^{-3w_0^2 + w_1^2 - 2w_2^2 + 0.4 w_0 w_1 w_2}$$

$$g(w_0, w_1, w_2) = \log(w_0^2 + w_1^2 + w_2^2 - 2w_1 w_2) - 3w_0^2 + w_1^2 - 2w_2^2 + 0.4 w_0 w_1 w_2$$

$$g(w_0, w_1, \ldots, w_n) = \frac{1}{1 + e^{-(w_0 + 3.27 w_1 + \ldots - 5.91 x_n)}}$$

$$g(w_0, w_1, \ldots, w_n) = \log\left(\frac{1}{1 + e^{-(w_0 + 3.27 w_1 + \ldots - 5.91 x_n)}}\right) + \log\left(\frac{1}{1 + e^{-(w_0 - 1.29 w_1 + \ldots + 0.22 x_n)}}\right) + \log\left(1 - \frac{1}{1 + e^{-(w_0 - 2.01 w_1 + \ldots + 0.73 x_n)}}\right)$$

Como achar o ponto de máximo da verossimilhança L(w)?

$$\hat{\mathbf{w}} = \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \\ \vdots \\ \hat{w}_n \end{bmatrix} = \arg\max_{\mathbf{w}} L(\mathbf{w})$$

- Equação de iteração de Newton uni-dimensional:

$$w^{k+1} = w^k - \frac{L'(w^k)}{L''(w^k)} = w^k - \left[L''(w^k)\right]^{-1} L'(w^k)$$

- Caso multivariado: a mesma coisa, apenas matricial

Como achar o ponto de máximo da verossimilhança L(w)?

- Equação de iteração de Newton uni-dimensional:

$$w^{k+1} = w^k - \frac{L'(w^k)}{L''(w^k)} = w^k - \left[L''(w^k)\right]^{-1} L'(w^k)$$

- Caso multivariado: a mesma coisa, apenas matricial

$$\mathbf{w}^{k+1} = \begin{bmatrix} w_0^{k+1} \\ w_1^{k+1} \\ \vdots \\ w_n^{k+1} \end{bmatrix} = \begin{bmatrix} w_0^{k} \\ w_1^{k} \\ \vdots \\ w_n^{k} \end{bmatrix} - \left[\underbrace{H(\mathbf{w}^k)}_{\text{matriz derivadas parciais de 2a ordem}} \right]^{-1} \underbrace{\nabla L(\mathbf{w}^k)}_{\text{vetor de derivadas parciais}}$$

- Para atualizar w_j usamos ***<u>TODAS</u>*** as derivadas parciais (com respeito a todos os w_p, a menos que H seja matriz diagonal), em contraste com métodos de gradiente

Relembre o modelo de regressão logística

- Dados: pares de vetores (x_i, y_i)
- x_i = idade da criança i
- y_i = 0 ou 1
- Cada criança joga uma moeda para determinar seu sucesso ou fracasso (Y_i)
- A probabilidade de sucesso da criança i depende de sua idade x_i

$$P(Y_i = 1 | x_i) = p(x_i) = \frac{1}{1+\exp(-(w_0 + w_1 x_i))}$$

- Resultados das crianças são independentes: produto das probabilidades individuais
- Qual a probabilidade de vermos os dados que temos?

A função de log-verossimilhança

- Com m crianças:

$$\begin{aligned} L(\mathbf{w}) &= L(w_0, w_1) \\ &= \mathbb{P}(Y_1 = 0, Y_2 = 1, \ldots, Y_m = 1) \\ &= \mathbb{P}(Y_1 = 0)\, \mathbb{P}(Y_2 = 1)\, \ldots\, \mathbb{P}(Y_m = 1) \\ &= \prod_{i=1}^{m} \mathbb{P}(Y_i = y_i) \end{aligned}$$

- onde cada y_i (minúsculo) é igual a 0 ou 1
- Temos $\mathbb{P}(Y_i = y_i) = \begin{cases} \sigma(x_i) & \text{se } y_i = 1 \\ 1 - \sigma(x_i) & \text{se } y_i = 0 \end{cases}$
- e $\sigma(x_i) = \frac{1}{1+e^{-(w_0+w_1 x_i)}}$

Um truque importante

- Vimos que $\mathbb{P}(Y_i = y_i) = \begin{cases} \sigma(x_i) & \text{se } y_i = 1 \\ 1 - \sigma(x_i) & \text{se } y_i = 0 \end{cases}$

- Podemos escrever esta expressão usando uma única linha:

$$\mathbb{P}(Y_i = y_i) = \sigma(x_i)^{y_i} (1 - \sigma(x_i))^{1-y_i}$$

- Você vai verificar isto na aula de exercícios

- Qual a vantagem? Tome log:

$$\log(\mathbb{P}(Y_i = y_i)) = y_i \ \log(\sigma(x_i)) + (1 - y_i) \ \log(1 - \sigma(x_i))$$

Log-verossimilhança

- Voltando para a amostra com os m indivíduos, obter a LOG-verossimilhança:

$$
\begin{aligned}
\ell(\mathbf{w}) &= \log(\ L(w_0, w_1)\ ) \\
&= \log\left(\prod_{i=1}^{m} \mathbb{P}(Y_i = y_i)\right) \\
&= \log\left(\prod_{i=1}^{m} \sigma(x_i)^{y_i} (1 - \sigma(x_i))^{1-y_i}\right) \\
&= \sum_{i=1}^{m} \log\left(\sigma(x_i)^{y_i} (1 - \sigma(x_i))^{1-y_i}\right) \\
&= \sum_{i=1}^{m} \left(y_i \ \log(\sigma(x_i)) + (1 - y_i)\ \log(1 - \sigma(x_i))\right)
\end{aligned}
$$

- sendo que $\sigma(x_i) = \frac{1}{1+e^{-(w_0+w_1 x_i)}}$

Equação de Newton: gradiente

- Precisamos das derivadas parciais com relação a w_0 e w_1
- Com $\sigma(x_i) = \sigma_i = 1/(1 + e^{-(w_0 + w_1 x_i)})$

- temos

$$\nabla \ell(\mathbf{w}) = \begin{bmatrix} \frac{\partial \log L}{\partial w_0} \\ \frac{\partial \log L}{\partial w_1} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} (y_i - \sigma_i) \\ \sum_{i=1}^{m} (y_i - \sigma_i) x_i \end{bmatrix} = m \begin{bmatrix} \bar{y} - \bar{\sigma} \\ \overline{xy} - \overline{\sigma x} \end{bmatrix} = m \begin{bmatrix} \overline{y - \sigma} \\ \overline{(y - \sigma) x} \end{bmatrix}$$

- onde os $\sigma(x_i) = \sigma_i$ são avaliados (calculados) com o valor corrente dos pesos e são médias aritméticas $\bar{y}$, $\bar{\sigma}$, $\overline{xy}$ e $\overline{\sigma x}$

Dedução passo a passo do gradiente

$$
\begin{aligned}
\frac{\partial \log L}{\partial w_0} &= \frac{\partial}{\partial w_0}\left[\sum_{i=1}^{m} y_i \log(\sigma(x_i)) + (1-y_i)\log(1-\sigma(x_i))\right] \\
&= \sum_{i=1}^{m}\left[y_i \frac{\partial \log(\sigma(x_i))}{\partial w_0} + (1-y_i)\frac{\partial \log(1-\sigma(x_i))}{\partial w_0}\right] \\
&= \sum_{i=1}^{m}\left[y_i \frac{1}{\sigma(x_i)}\frac{\partial \sigma(x_i)}{\partial w_0} + (1-y_i)\frac{1}{1-\sigma(x_i)}\frac{\partial(-\sigma(x_i))}{\partial w_0}\right] \\
&= \sum_{i=1}^{m}\left[\frac{y_i}{\sigma(x_i)} - \frac{1-y_i}{1-\sigma(x_i)}\right]\left[\frac{\partial \sigma(x_i)}{\partial w_0}\right]
\end{aligned}
$$

Dedução passo a passo do gradiente

$$
\begin{aligned}
\frac{\partial \sigma(x_i)}{\partial w_0} &= \frac{\partial}{\partial w_0} \frac{1}{1+e^{-(w_0+w_1 x_i)}} \\
&= \frac{\partial}{\partial w_0} \left(1+e^{-(w_0+w_1 x_i)}\right)^{-1} \\
&= (-1)\left(1+e^{-(w_0+w_1 x_i)}\right)^{-2} \frac{\partial e^{-(w_0+w_1 x_i)}}{\partial w_0} \\
&= \frac{-1}{(1+e^{-(w_0+w_1 x_i)})^2} e^{-(w_0+w_1 x_i)} \frac{\partial(-(w_0+w_1 x_i))}{\partial w_0} \\
&= \frac{-1}{(1+e^{-(w_0+w_1 x_i)})^2} e^{-(w_0+w_1 x_i)} (-1) \\
&= \frac{1}{1+e^{-(w_0+w_1 x_i)}} \frac{e^{-(w_0+w_1 x_i)}}{1+e^{-(w_0+w_1 x_i)}} \\
&= \sigma(x_i)(1-\sigma(x_i))
\end{aligned}
$$

Dedução passo a passo do gradiente

$$\frac{\partial \log L}{\partial w_0} = \sum_{i=1}^{m} \left[\frac{y_i}{\sigma(x_i)} - \frac{1-y_i}{1-\sigma(x_i)} \right] \left[\sigma(x_i)(1-\sigma(x_i))\right]$$

$$= \sum_{i=1}^{m} \left[y_i(1-\sigma(x_i)) - (1-y_i)\sigma(x_i)\right]$$

$$= \sum_{i=1}^{m} \left[y_i - y_i\sigma(x_i)) - \sigma(x_i) + y_i\sigma(x_i)\right]$$

$$= \sum_{i=1}^{m} \left[y_i - \sigma(x_i)\right]$$

Dedução passo a passo do gradiente

$$\frac{\partial \log L}{\partial w_1} = \sum_{i=1}^{m} \left[\frac{y_i}{\sigma(x_i)} - \frac{1-y_i}{1-\sigma(x_i)} \right] \left[\frac{\partial \sigma(x_i)}{\partial w_1} \right]$$

$$\begin{aligned}
\frac{\partial \sigma(x_i)}{\partial w_1} &= \frac{\partial}{\partial w_0} \frac{1}{1+e^{-(w_0+w_1 x_i)}} \\
&= \frac{-1}{\left(1+e^{-(w_0+w_1 x_i)}\right)^{-2}} e^{-(w_0+w_1 x_i)} \frac{\partial(-(w_0+w_1 x_i))}{\partial w_1} \\
&= -\sigma(x_i)(1-\sigma(x_i))\,(-x_i) \\
&= \sigma(x_i)(1-\sigma(x_i))\,x_i
\end{aligned}$$

Dedução passo a passo do gradiente

$$\frac{\partial \log L}{\partial w_1} = \sum_{i=1}^{m} \left[\frac{y_i}{\sigma(x_i)} - \frac{1-y_i}{1-\sigma(x_i)} \right] \left[\frac{\partial \sigma(x_i)}{\partial w_1} \right]$$

$$= \sum_{i=1}^{m} \left[\frac{y_i}{\sigma(x_i)} - \frac{1-y_i}{1-\sigma(x_i)} \right] \left[\sigma(x_i)(1-\sigma(x_i))\, x_i \right]$$

$$= \sum_{i=1}^{m} \left[y_i - \sigma(x_i) \right] (x_i)$$

Vetorizando o gradiente

$$\frac{\partial \ell}{\partial \mathbf{w}} = \underbrace{\frac{\partial \log L}{\partial \mathbf{w}}}_{2\times 1}$$

$$= \begin{bmatrix} \partial\ell/\partial w_0 \\ \partial\ell/\partial w_1 \end{bmatrix} = \begin{bmatrix} \sum_i (y_i - \sigma(x_i))\, 1 \\ \sum_i (y_i - \sigma(x_i))\, x_i \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_m \end{bmatrix} \begin{bmatrix} y_1 - \sigma(x_1) \\ y_2 - \sigma(x_2) \\ \vdots \\ y_m - \sigma(x_m) \end{bmatrix}$$

Equação de Newton: Hessiano

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial w_0^2} &= \frac{\partial}{\partial w_0} \frac{\partial \ell}{\partial w_0} \\
&= \frac{\partial}{\partial w_0} \left(\sum_i [y_i - \sigma(x_i)] \right) \\
&= \sum_i \frac{\partial}{\partial w_0} [y_i - \sigma(x_i)] = -\sum_i \frac{\partial \sigma(x_i)}{\partial w_0} \\
&= -\sum_i \sigma(x_i)(1 - \sigma(x_i)) = -n \frac{1}{n} \sum_i \sigma(x_i)(1 - \sigma(x_i)) \\
&= -n \overline{\sigma(1 - \sigma)}
\end{aligned}
$$

Dedução passo a passo do Hessiano

- De modo similar, obtemos os demais elementos da matriz Hessiana.

$$H = -n \begin{bmatrix} \overline{\sigma(1-\sigma)} & \overline{\sigma(1-\sigma)x} \\ \sigma(1-\sigma)x & \sigma(1-\sigma)x^2 \end{bmatrix}$$

- onde os elementos acima são médias aritméticas sobre os exemplos

Equação de iteração de Newton

- De volta ao procedimento de maximização:

$$\mathbf{w}^{k+1} = \begin{bmatrix} w_0^{k+1} \\ w_1^{k+1} \end{bmatrix} = \begin{bmatrix} w_0^{k} \\ w_1^{k} \end{bmatrix} - \left[\underbrace{H(\mathbf{w}^k)}_{\text{matriz derivadas parciais de 2a ordem}} \right]^{-1} \underbrace{\nabla L(\mathbf{w}^k)}_{\text{vetor de derivadas parciais}}$$

$$\mathbf{w}^{k+1} = \begin{bmatrix} w_0^{k+1} \\ w_1^{k+1} \end{bmatrix} = \begin{bmatrix} w_0^{k} \\ w_1^{k} \end{bmatrix} + \frac{1}{n} \begin{bmatrix} \overline{\sigma(1-\sigma)} & \overline{\sigma(1-\sigma)x} \\ \overline{\sigma(1-\sigma)x} & \overline{\sigma(1-\sigma)x^2} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_m \end{bmatrix} \begin{bmatrix} y_1 - \sigma(x_1) \\ y_2 - \sigma(x_2) \\ \vdots \\ y_m - \sigma(x_m) \end{bmatrix}$$

- Para atualizar w_1, usamos a derivada parcial com respeito a w_1 E TAMBÉM w_0 (a menos que H seja matriz diagonal, e geralmente ela não é diagonal).

Flexibilidade da regressão logística

- Regressão logística é menos limitada do que parece.
- Os inputs-features podem ser:
  - Quaisquer características (features) dos dados
  - Transformações das features x originais tais como, por exemplo, log(x)
  - Uma expansão de base, por exemplo, x**2 e x**3
  - Indicadores de categorias (features categóricas)
  - Interações entre duas features tal como, por exemplo, x2 * x3
- A simplicidade e flexibilidade da regressão logística a tornam uma das técnicas de classificação estatística mais importantes e mais amplamente utilizada.

Regressão logística com várias features

- A chance de sucesso da criança não depende APENAS de sua idade.
- Vai depender também de:
  - sexo: feature $X_2 = 0$ (masc) ou 1 (fem)
  - escolaridade da mãe: feature X_3 = no. de anos de estudo formal
  - renda per capita da família: feature X_4 = renda mensal em 1000 reais
- Coletamos as features de cada criança num vetor **x** (em negrito):
  - $\mathbf{x} = (x_1, x_2, x_3, x_4)$
- Como fazer um modelo em que a chance de sucesso depende de todas estas características simultaneamente?

- Modelo logístico incorpora todas as features de forma LINEAR.
- Para cada criança, crie um escore z:
  - Cada feature da criança é multiplicada por um peso w
  - O peso da feature está associado à importância da feature:
    - features importantes terão |w| grande
    - features pouco importantes terão seu peso |w| pequeno
    - features totalmente irrelevantes devem ter |w| aprox zero
  - Depois de ponderar cada feature da criança, somamos para obter o escore z

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4$$

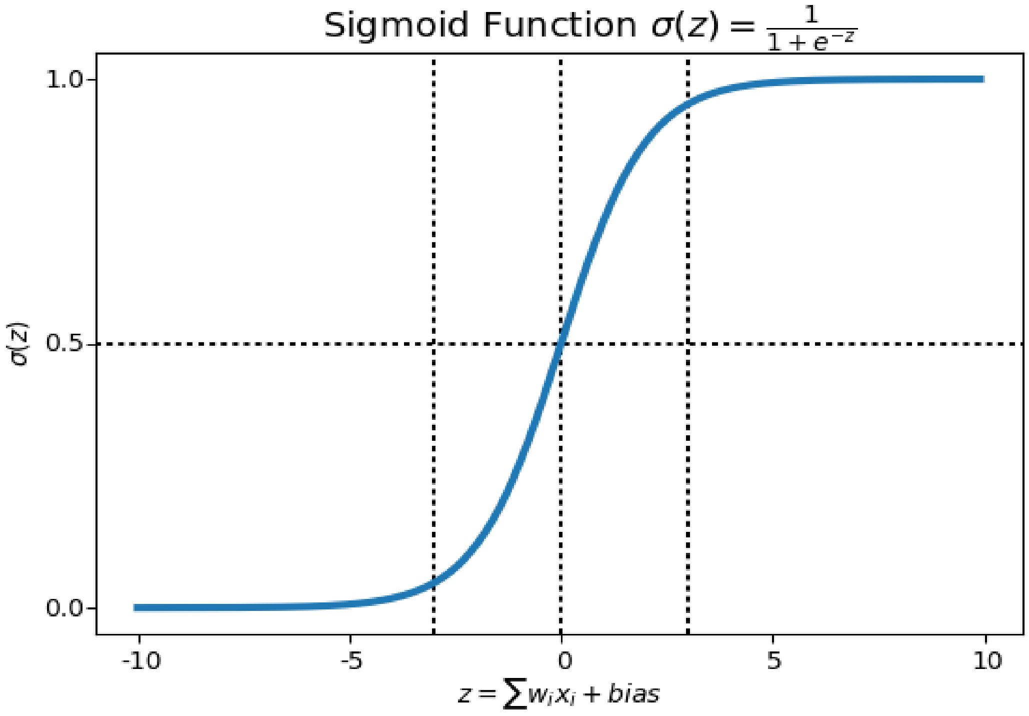
Regressão logística com várias features

- Calcule $z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4$
- para cada criança
- Queremos que a probabilidade de sucesso seja uma função do escore:
    - um alto valor de z leva a uma probabilidade alta (aprox 1)
    - um valor baixo de z leva a uma probabilidade baixa (aprox 0)
- Reduzimos a complexidade da análise a uma forma manejável, simples.
- O escore z embute a influência de todas as features ao mesmo tempo.

$$\mathbb{P}(Y = 1) = \sigma(z) = \frac{1}{1+e^{-z}}$$

- Dois indivíduos com features diferentes *MAS COM O MESMO ESCORE z* terão a mesma probabilidade de sucesso.

Representação gráfica



$$\mathbb{P}(Y = 1) = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4$$

Aprendizagem a partir dos dados

- Precisamos responder várias perguntas:
  - 1) Este modelo logístico representa bem os dados observados?
  - 2) Se ele representar bem os dados, como aprender os pesos "corretos" a partir de dados observados (= amostra de treinamento)
  - 3) Não queremos apenas aprender com os dados. Queremos a "melhor representação" possível. Qual a "melhor maneira" de aprender os pesos?
  - 4) Podemos fazer algo melhor que usar a regressão logística?
- Vamos responder (2) e (3) no resto dessa aula. Amanhã, veremos (1) e (4).

Olhando os escores de toda a amostra de treinamento

- Imagine que temos n=4 features e m crianças.
- Calculamos os escores z de todas elas numa única operação matricial:

$$\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} \\ & & \vdots & & \\ & & \vdots & & \\ 1 & x_{m1} & x_{m2} & x_{m3} & x_{m4} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} \mathbf{w}'\mathbf{x}^{(1)} \\ \mathbf{w}'\mathbf{x}^{(2)} \\ \vdots \\ \vdots \\ \mathbf{w}'\mathbf{x}^{(m)} \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ \vdots \\ z_m \end{bmatrix}$$

- ***<u>Um único</u>*** vetor de pesos w é aplicado a cada uma das m crianças

Calcule agora as probabilidades de sucesso

- Depois de obter os z's obtenha as probabilidades:

$$\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} \\ & & \vdots & & \\ & & \vdots & & \\ 1 & x_{m1} & x_{m2} & x_{m3} & x_{m4} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} \mathbf{w}'\mathbf{x}^{(1)} \\ \mathbf{w}'\mathbf{x}^{(2)} \\ \vdots \\ \vdots \\ \mathbf{w}'\mathbf{x}^{(m)} \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ \vdots \\ z_m \end{bmatrix} \rightarrow \begin{bmatrix} \sigma(z_1) \\ \sigma(z_2) \\ \vdots \\ \vdots \\ \sigma(z_m) \end{bmatrix} = \begin{bmatrix} 1/(1+e^{-z_1}) \\ 1/(1+e^{-z_2}) \\ \vdots \\ \vdots \\ 1/(1+e^{-z_m}) \end{bmatrix}$$

- Como obter os pesos w?
    - Do mesmo modo que antes: maximize a log-verossimilhança
    - Fórmulas são as mesmas de antes

Equação de iteração de Newton

- De volta ao procedimento de maximização:

$$\mathbf{w}^{k+1} = \begin{bmatrix} w_0^{k+1} \\ w_1^{k+1} \end{bmatrix} = \begin{bmatrix} w_0^{k} \\ w_1^{k} \end{bmatrix} - \Bigg[\underbrace{H(\mathbf{w}^k)}_{\text{matriz derivadas parciais de 2a ordem}} \Bigg]^{-1} \underbrace{\nabla L(\mathbf{w}^k)}_{\text{vetor de derivadas parciais}}$$

$$\mathbf{w}^{k+1} = \begin{bmatrix} w_0^{k+1} \\ w_1^{k+1} \end{bmatrix} = \begin{bmatrix} w_0^{k} \\ w_1^{k} \end{bmatrix} + \frac{1}{m} \begin{bmatrix} \overline{\sigma(1-\sigma)} & \overline{\sigma(1-\sigma)x} \\ \overline{\sigma(1-\sigma)x} & \overline{\sigma(1-\sigma)x^2} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_m \end{bmatrix} \begin{bmatrix} y_1 - \sigma(x_1) \\ y_2 - \sigma(x_2) \\ \vdots \\ y_m - \sigma(x_m) \end{bmatrix}$$

- Para atualizar w_1, usamos a derivada parcial com respeito a w_1 E TAMBÉM w_0 (a menos que H seja matriz diagonal, e geralmente ela não é diagonal).

Log-verossimilhança

$$\begin{aligned}
\ell(\mathbf{w}) &= \log(\ L(w_0, w_1, w_2, w_3, w_4)\ ) \\
&= \log\left(\prod_{i=1}^{m} \mathbb{P}(Y_i = y_i)\right) \\
&= \log\left(\prod_{i=1}^{m} \sigma(\mathbf{x}^{(i)})^{y_i}\,(1-\sigma(\mathbf{x}^{(i)}))^{1-y_i}\right) \\
&= \sum_{i=1}^{m} \log\left(\sigma(\mathbf{x}^{(i)})^{y_i}\,(1-\sigma(\mathbf{x}^{(i)}))^{1-y_i}\right) \\
&= \sum_{i=1}^{m} \left(y_i\ \log(\sigma(\mathbf{x}^{(i)})) + (1-y_i)\ \log(1-\sigma(\mathbf{x}^{(i)}))\right)
\end{aligned}$$

- sendo que $\sigma(\mathbf{x}^{(i)}) = \frac{1}{1+e^{-z_i}} = \frac{1}{1+e^{-(w_0+w_1x_{i1}+w_2x_{i2}+w_3x_{i3}+w_4x_{i4})}}$

Equação de Newton: gradiente

- Precisamos das derivadas parciais com relação a cada componente de w
- Temos

$$\nabla \ell(\mathbf{w}) = \begin{bmatrix} \frac{\partial \log L}{\partial w_0} \\ \frac{\partial \log L}{\partial w_1} \\ \frac{\partial \log L}{\partial w_2} \\ \frac{\partial \log L}{\partial w_3} \\ \frac{\partial \log L}{\partial w_4} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}(y_i - \sigma_i) \\ \sum_{i=1}^{n}(y_i - \sigma_i)x_{i1} \\ \sum_{i=1}^{n}(y_i - \sigma_i)x_{i2} \\ \sum_{i=1}^{n}(y_i - \sigma_i)x_{i3} \\ \sum_{i=1}^{n}(y_i - \sigma_i)x_{i4} \end{bmatrix} = n \begin{bmatrix} \overline{y - \sigma} \\ \overline{(y - \sigma)x_1} \\ \overline{(y - \sigma)x_2} \\ \overline{(y - \sigma)x_3} \\ \overline{(y - \sigma)x_4} \end{bmatrix}$$

Mais uma forma de expressar o gradiente

- Notação matricial

$$\nabla \ell(\mathbf{w}) = \begin{bmatrix} \frac{\partial \log L}{\partial w_0} \\ \frac{\partial \log L}{\partial w_1} \\ \frac{\partial \log L}{\partial w_2} \\ \frac{\partial \log L}{\partial w_3} \\ \frac{\partial \log L}{\partial w_4} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} (y_i - \sigma_i) \\ \sum_{i=1}^{n} (y_i - \sigma_i) x_{i1} \\ \sum_{i=1}^{n} (y_i - \sigma_i) x_{i2} \\ \sum_{i=1}^{n} (y_i - \sigma_i) x_{i3} \\ \sum_{i=1}^{n} (y_i - \sigma_i) x_{i4} \end{bmatrix} = \underbrace{\begin{bmatrix} | & | & & & | \\ \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \ldots & \ldots & \mathbf{x}^{(m)} \\ | & | & & & | \end{bmatrix}}_{5 \times m} \underbrace{\begin{bmatrix} y_1 - \sigma_1 \\ y_2 - \sigma_2 \\ \vdots \\ y_m - \sigma_m \end{bmatrix}}_{m \times 1}$$

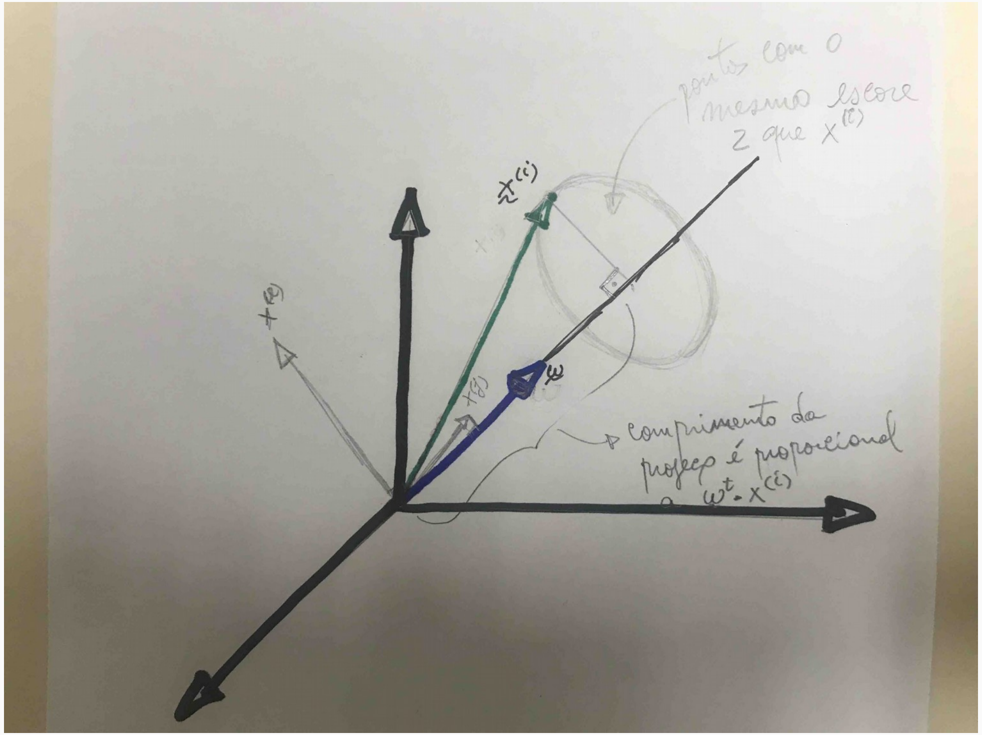
Equação de iteração de Newton

- De volta ao procedimento de maximização:

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \left[\underbrace{H(\mathbf{w}^k)}_{\text{5 x 5, 2a ordem}} \right]^{-1} \underbrace{\nabla L(\mathbf{w}^k)}_{\text{vetor 5 x 1}}$$

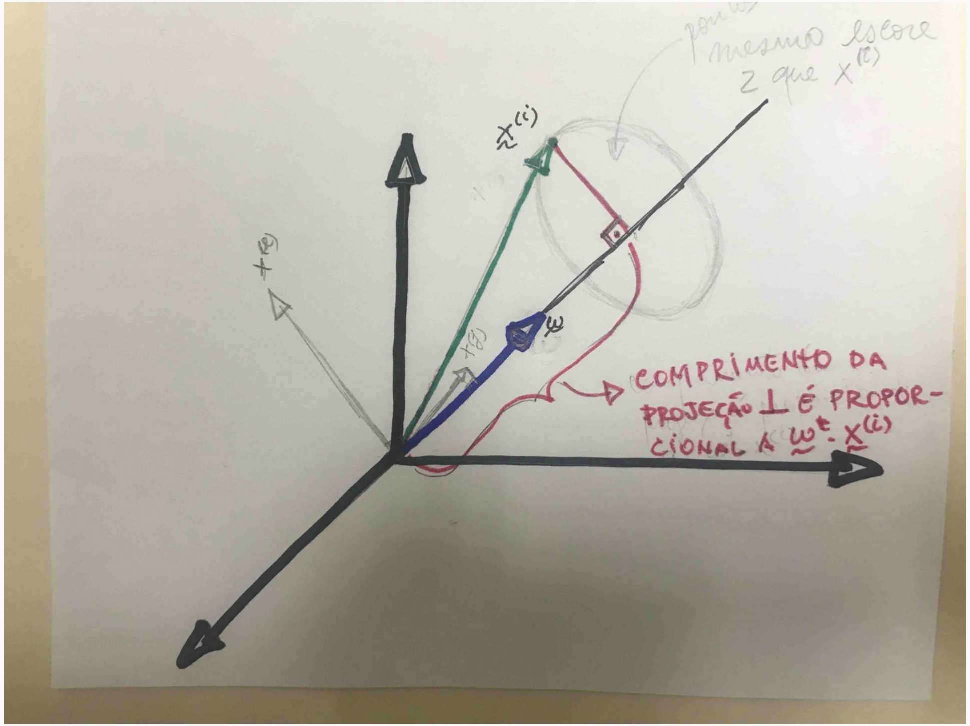
$$\mathbf{w}^{k+1} = \mathbf{w}^k + \frac{1}{m} \begin{bmatrix} \overline{\sigma(1-\sigma)} & \overline{\sigma(1-\sigma)x_1} & \dots & \overline{\sigma(1-\sigma)x_4} \\ \overline{\sigma(1-\sigma)x_1} & \overline{\sigma(1-\sigma)x_2^2} & \dots & \overline{\sigma(1-\sigma)x_2x_4} \\ & & \vdots & \\ \overline{\sigma(1-\sigma)x_4} & \overline{\sigma(1-\sigma)x_1x_4} & \dots & \overline{\sigma(1-\sigma)x_4^2} \end{bmatrix}^{-1} \begin{bmatrix} | & | & & & | \\ \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \dots & \dots & \mathbf{x}^{(m)} \\ | & | & & & | \end{bmatrix} \begin{bmatrix} y_1 - \sigma_1 \\ y_2 - \sigma_2 \\ \vdots \\ y_m - \sigma_m \end{bmatrix}$$

- O que significa $z_i = \mathbf{w}'\mathbf{x}^{(i)} = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + w_4 x_{i4}$?



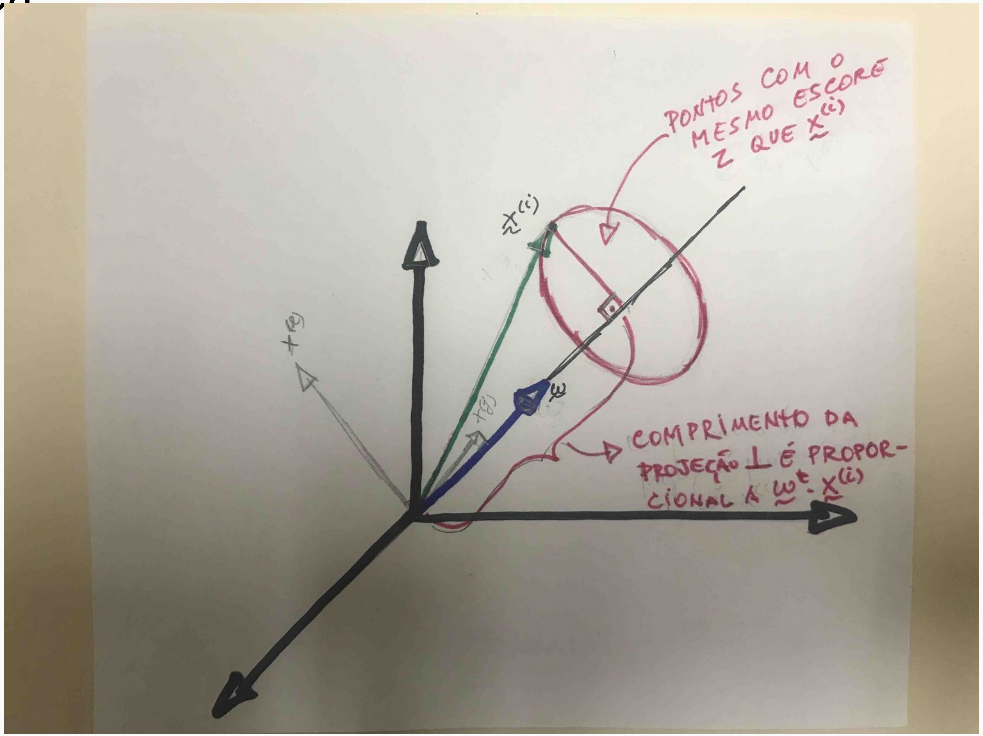
Representação geométrica

- O que significa $z_i = \mathbf{w}'\mathbf{x}^{(i)} = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + w_4 x_{i4}$?



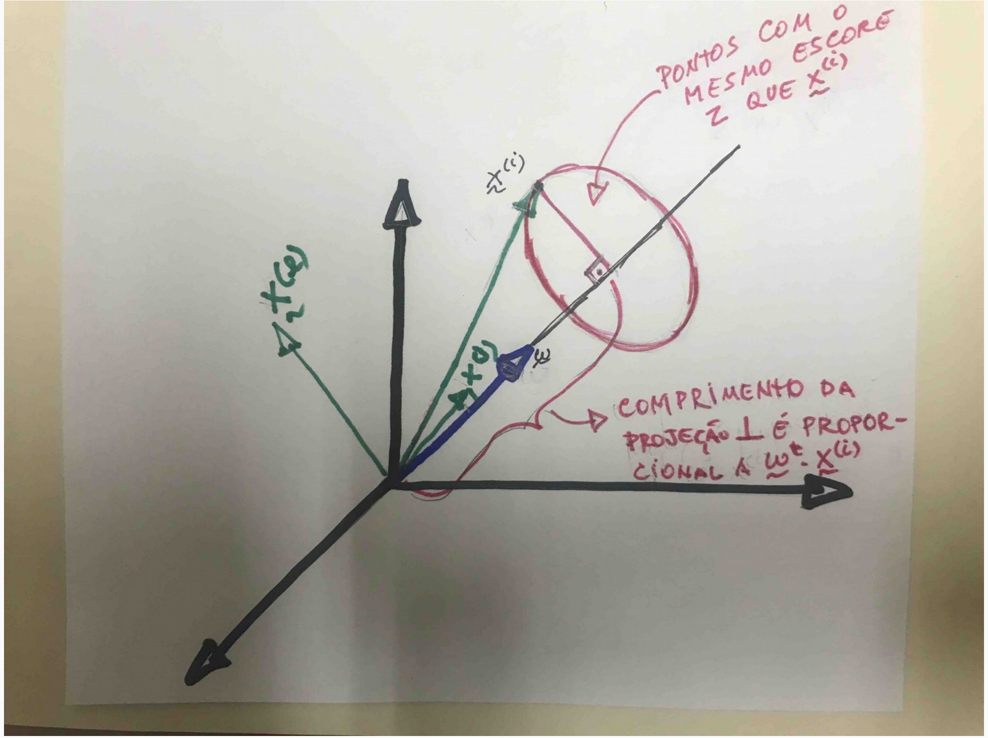
Representação geométrica

- O que significa $z_i = \mathbf{w}'\mathbf{x}^{(i)} = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + w_4 x_{i4}$?

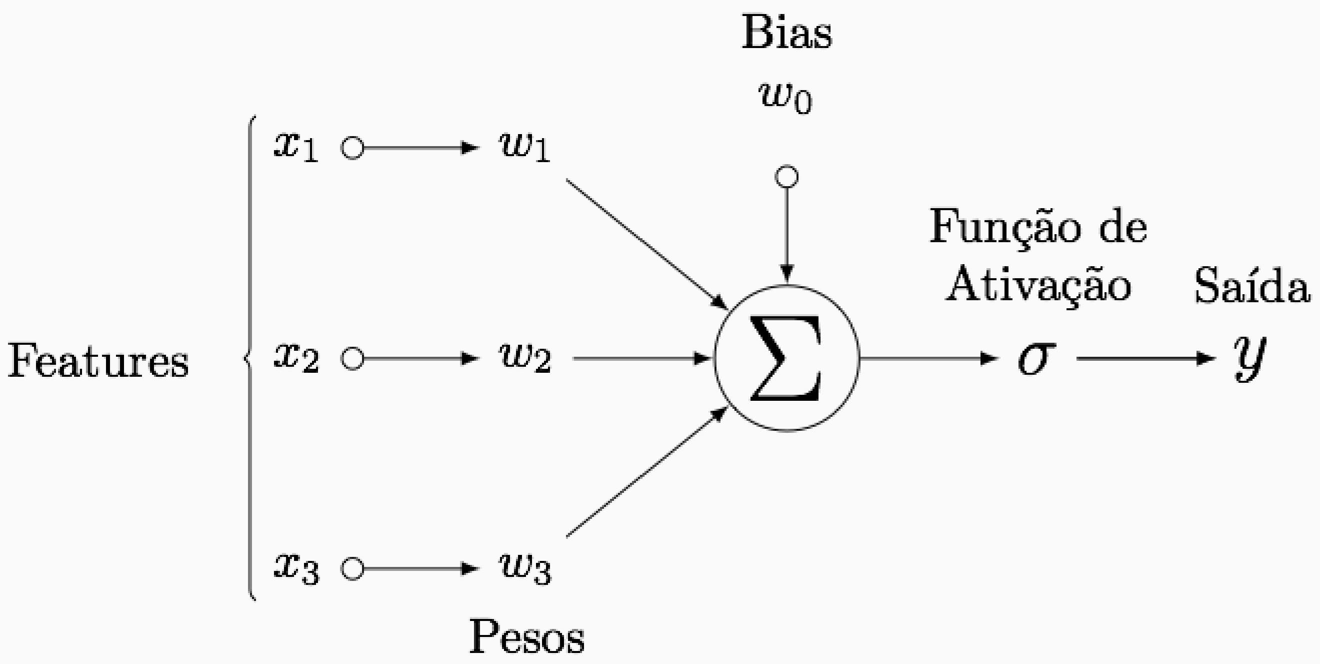


Representação geométrica

- O que significa $z_i = \mathbf{w}'\mathbf{x}^{(i)} = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + w_4 x_{i4}$?

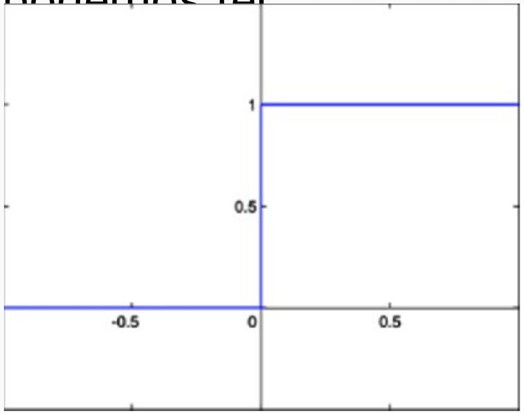


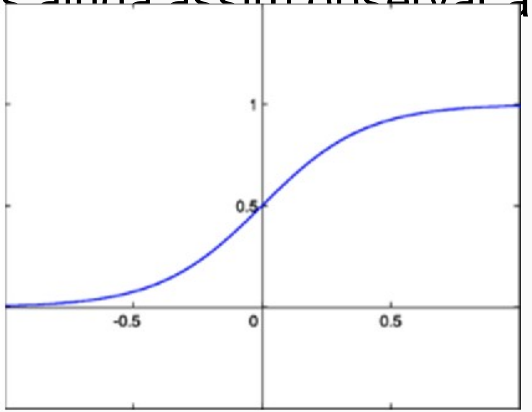
Regressão logística como rede neural com uma camada



Perceptron x logística

- Perceptron: threshold "hard": se $z = w_0 + w_1 x_1 + \ldots + w_n x_n > 0 \rightarrow \mathbb{P}(\text{ class1}) = 1$
- Modelo logístico: threshold "soft": se $z = w_0 + w_1 x_1 + \ldots + w_n x_n > 0 \rightarrow \mathbb{P}(\text{ class1}) > 1/2$
    - perceptron gera dados com classes linearmente separáveis
    - logística gera dados não-linearmente separáveis:
        - podemos ter $\mathbb{P}(\text{ class1}) \approx 1$ mas ainda assim observar a classe 0





Usando a regressão logística para classificar

- Imagine que temos apenas duas features, x_1 e x_2
- Achamos os pesos w_0, w_1 e w_2 por máxima verossimilhança
- Temos então

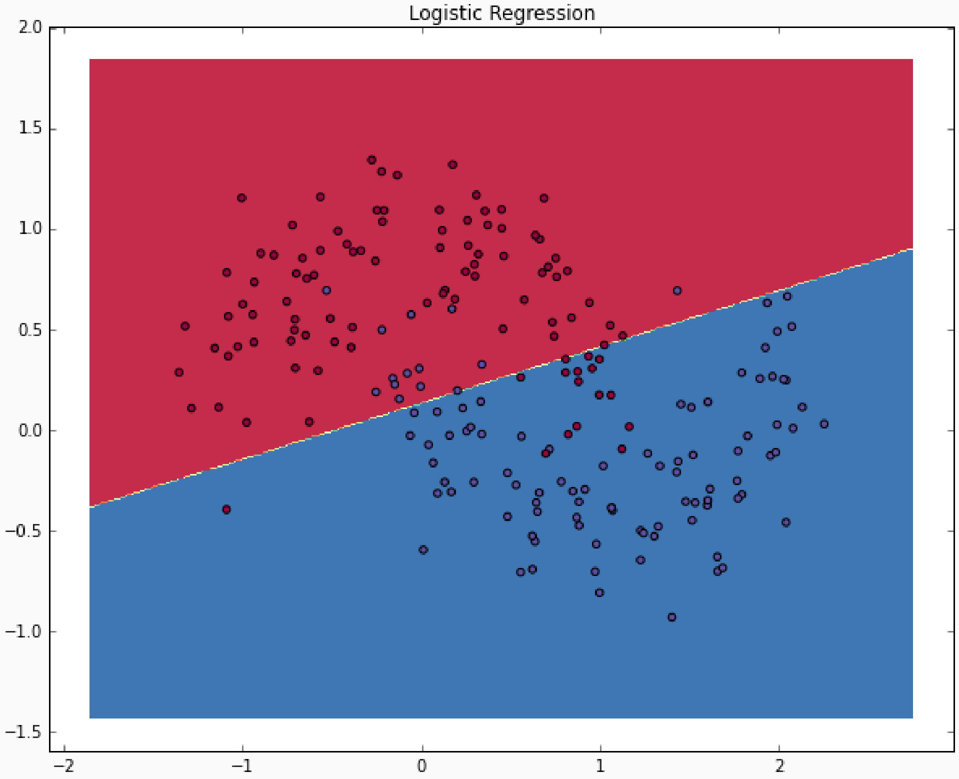
$$\mathbb{P}(Y = 1 | x_1, x_2) = \sigma(x_1, x_2) = \frac{1}{1+e^{w_0 + w_1 x_1 + w_2 x_2}}$$

- Considere os pontos do plano (x_1, x_2) tais que esta probab = ½
- Quem são estes pontos? (Exercício)
- São os pontos tais que

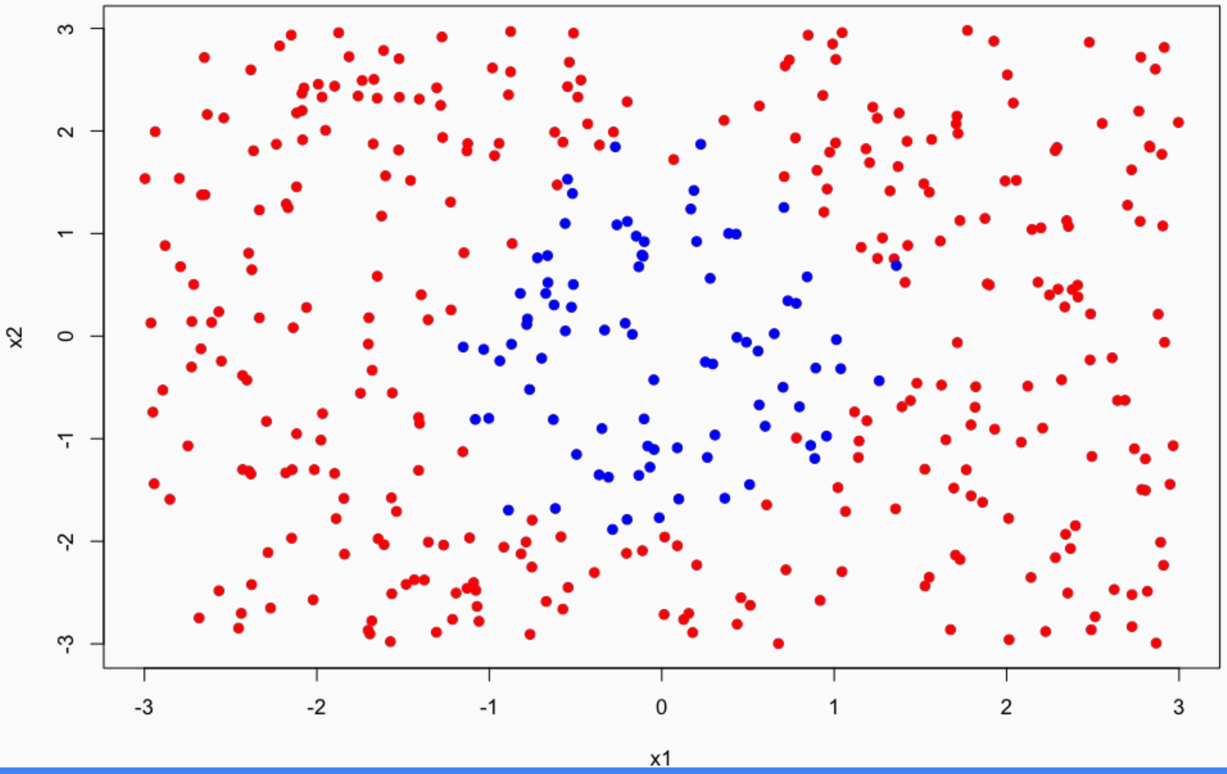
$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

- Esta é a equação de uma reta no plano (x_1, x_2)
- Ela determina uma fronteira de decisão:
  - de um lado, probab de sucesso é > ½
  - do outro lado, é menor que ½

Decision boundary



E quando a real fronteira de decisão não for linear?

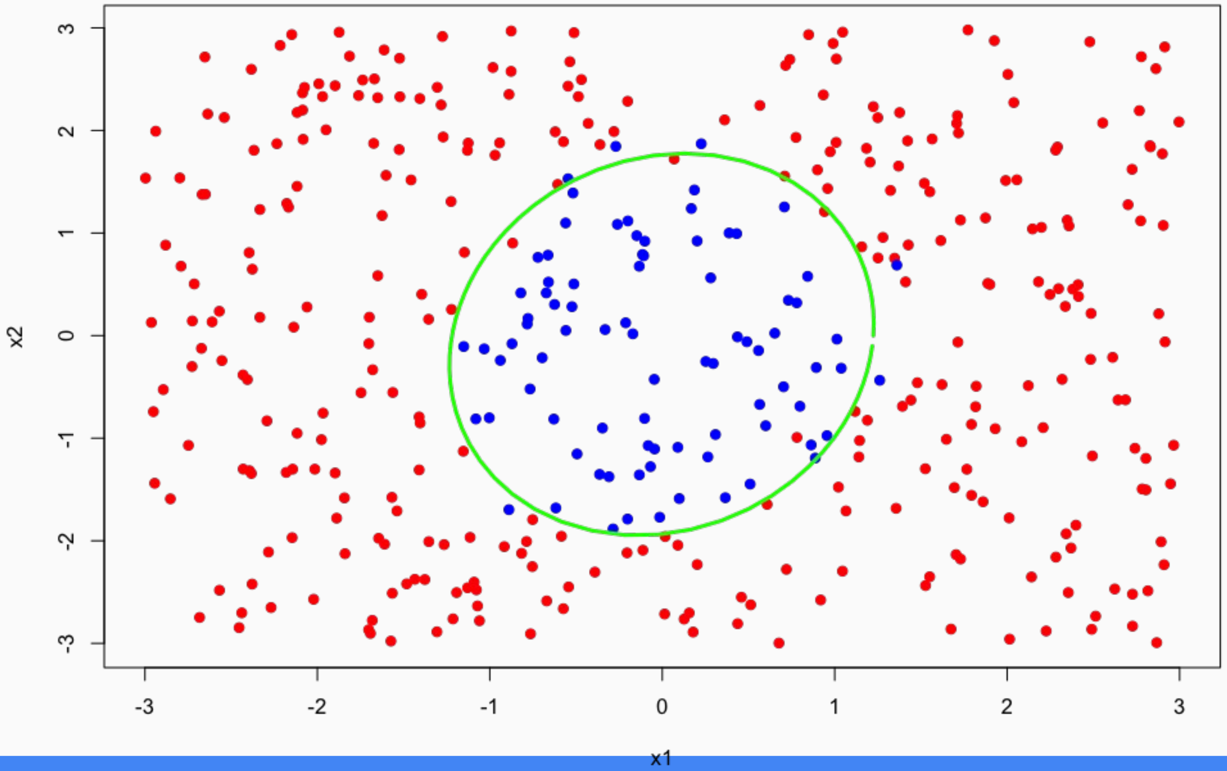


Modelo generativo usado e ajuste de regressão logística

$$z_i = 7 - 0.1x_{i1} - 0.15x_{i2} - 4.4x_{i1}^2 - 2.2x_{i2}^2 + 0.5x_{i1}x_{i2}$$

```
> summary(fit1)

Call:
glm(formula = y ~ matx, family = binomial("logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.34601  -0.00377   0.00000   0.00000   2.48558

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.59718    1.92892   4.975 6.51e-07 ***
matxx1       0.03001    0.42974   0.070   0.9443
matxx2      -0.47726    0.26839  -1.778   0.0754 .
matx        -6.43045    1.29979  -4.947 7.52e-07 ***
matx        -2.80773    0.56631  -4.958 7.13e-07 ***
matx         0.93236    0.47525   1.962   0.0498 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 411.165  on 399  degrees of freedom
Residual deviance:  54.943  on 394  degrees of freedom
AIC: 66.943

Number of Fisher Scoring iterations: 11
```

Resultado do ajuste: fronteira de decisão



Flexibilidade da regressão logística

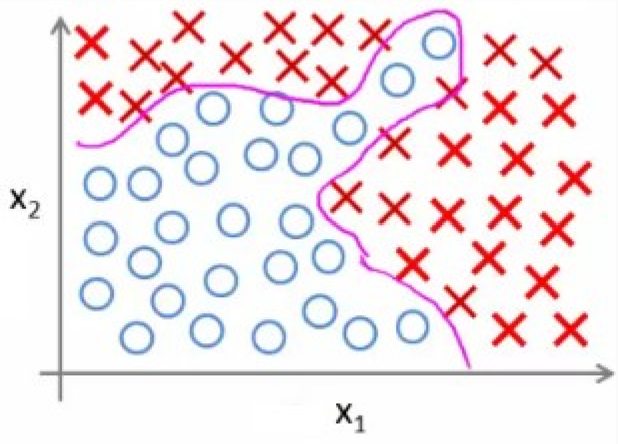
- Este exemplo mostra que a regressão logística possui grande flexibilidade
- Features podem ser criadas a partir de features básicas:
    - potências de features básicas: x2 = x1**2 (renda ao quadrado, ao cubo)
    - transformações não-lineares de features básicas: x2 = g(x1) (tal como log(renda) ou sqrt(renda))
    - termos de interações entre features: x3 = x1*x2 (tal como x3 = sexo*renda)
- A probabilidade de sucesso é uma função de uma COMBINAÇÃO LINEAR das features (básicas ou derivadas):

$$z_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i1}^2 + w_4 x_{i1} x_{i2} + w_5 \log(x_{i1}) + w_6 x_{i2} \sqrt{x_{i3}}$$

$$\mathbb{P}(Y_i = 1 | \mathbf{x}_i) = \frac{1}{1+e^{-z_i}}$$

Importante mensagem:

- Para aprender uma decision boundary não-linear com regressão logística → precisamos de muitos termos não lineares das features "básicas"
- Por exemplo, com duas features x1 e x2, podemos buscar os pesos w com

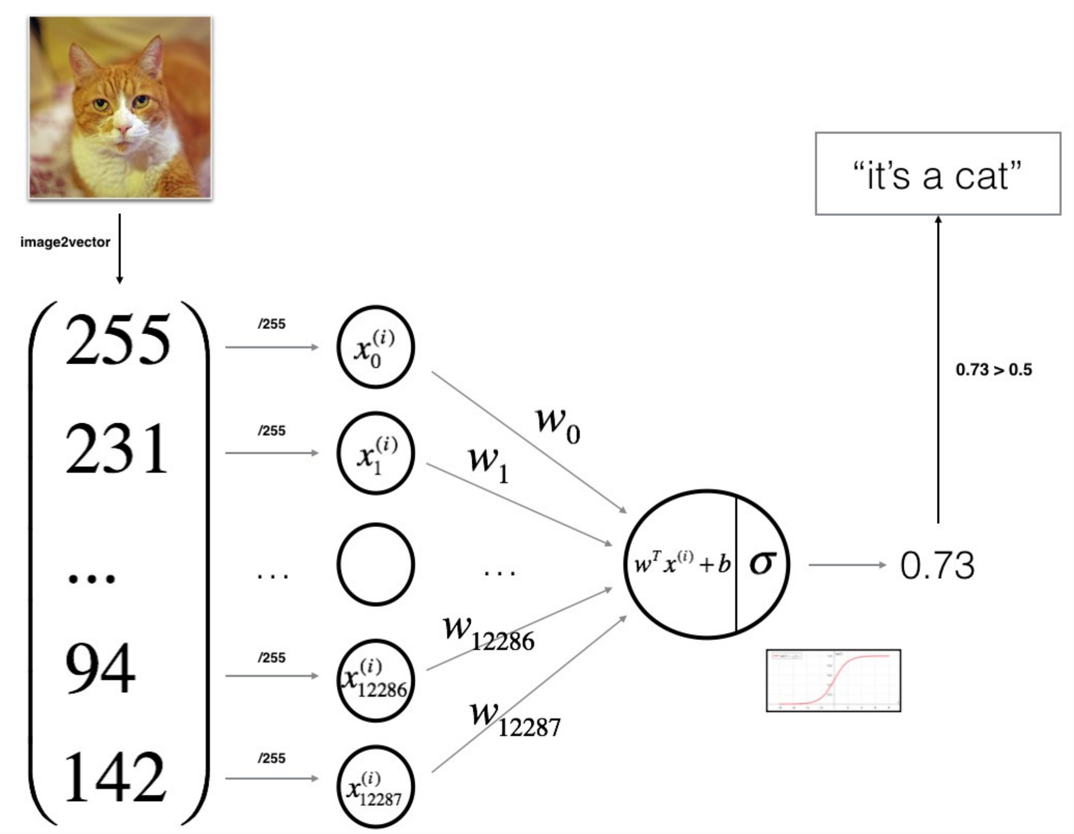


$$\mathbb{P}(Y = 1 | x_1, x_2) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2+w_3x_1^2+w_4x_2^2+w_5x_1x_2+w_6x_1^3+w_7x_1^2x_2+w_8x_1x_2^2)}}$$

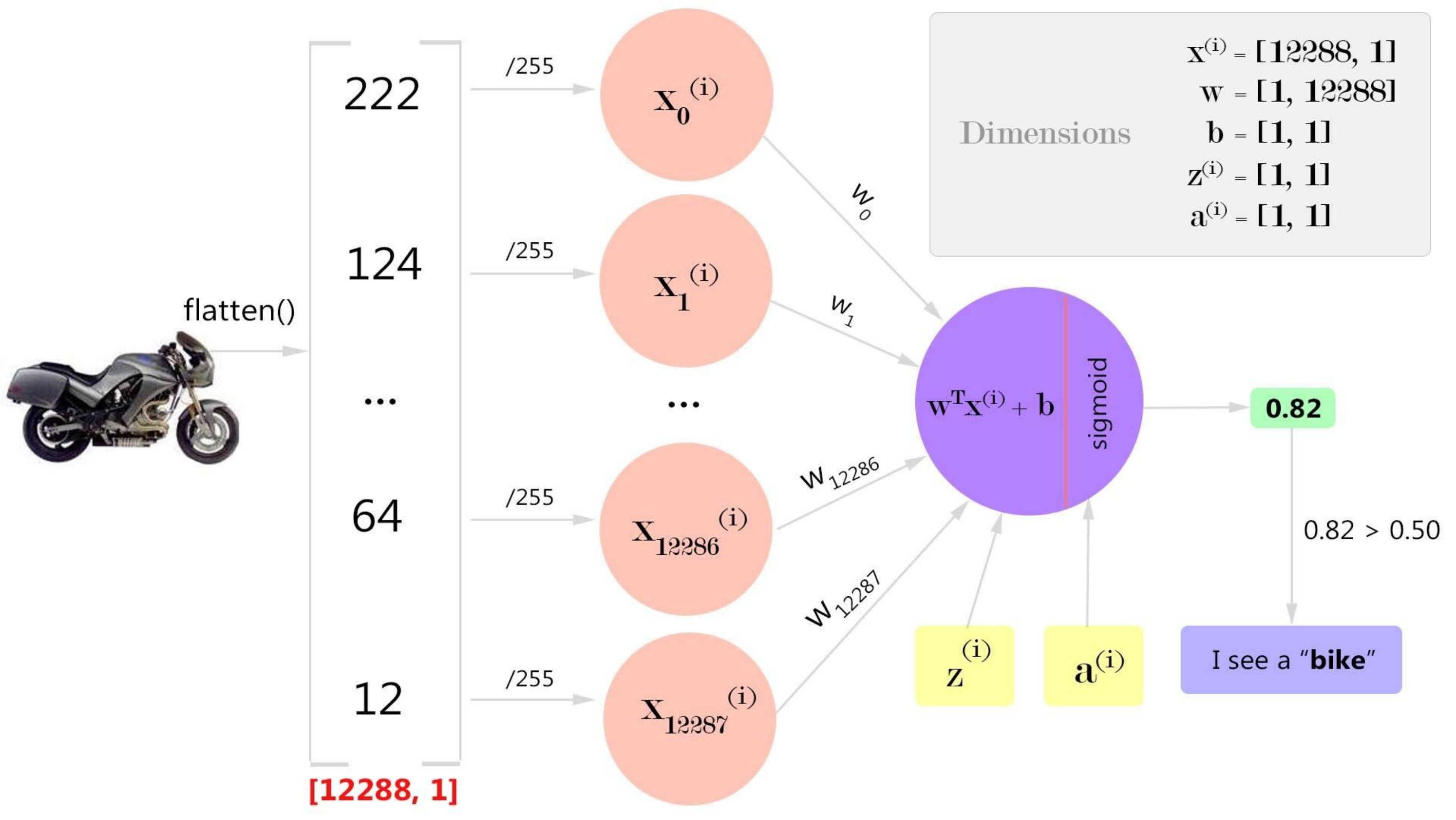
Regressão logística para imagem?

- Podemos usar regressão logística para classificar imagens em dois grupos.
- Por exemplo, gatos x não-gatos
- Provavelmente não teremos uma bom resultado
- Mas como isto pode ser feito, mesmo que gerando um resultado pobre em termos de acertos na classificação?
- Transformamos cada imagem num grande vetor de features.
- As features são as intensidades de "cores"nos pixels das imagens.
- Isto é,
  - cada pixel → uma feature.

Logística para imagem?



Logística para imagem?



Métricas para avaliar a regra de classificação

- A classificação feita pela nossa regra de decisão (baseda na regressão logística não é perfeita.
- Ela comete vários erros: indivíduoa que de fato são diabéticos não possuem as características x_1 e x_2 típicas de um diabético.
- Em consequência, a nossa regra de decisão (que olha apenas os regressores em $\boldsymbol{x}$) aloca estes indivíduos à classe 0 (não diabéticos).
- Estes são os *falso-negativos* (o diagnóstico é falsamente negativo).
- Analogamente, vários não-diabéticos possuem características típicas de diabéticos e são então alocados pela regra de decisão logística à categora 1 (diabéticos).
- Estes são os *falso-positivos* (o diagnóstico é falsamente positivo).
- Claro, existe o conceito de *verdadeiro-positivo* e *verdadeiro-negativo*.

Falso-positivos e Falso-negativos

- Idealmente, queremos poucos falso-positivos e poucos falso-negativos (ou muitos verdadeiro-positivos e muitos verdadeiro-negativos).
- Isto será obtido se tivermos uma pequena probabilidade de ter um falso-positivo (FP) e um falso-negativo (FN).

$$\mathbb{P}(FP) = \mathbb{P}(\text{classificado como } + | \text{é -}) = \frac{\mathbb{P}(\text{classif } + \text{ e é -})}{\mathbb{P}(\text{é -})}$$

e

$$\mathbb{P}(FN) = \mathbb{P}(\text{classificado como -} | \text{é } +) = \frac{\mathbb{P}(\text{classif - e é } +)}{\mathbb{P}(\text{é -})}$$

- No caso de verdadeiro-positivos , temos

$$\mathbb{P}(VP) = \mathbb{P}(\text{classificado como } + | \text{é } +) = \frac{\mathbb{P}(\text{classif } + \text{ e é } +)}{\mathbb{P}(\text{é } +)}$$

Recall ou revocação ou sensibilidade

- No caso de verdadeiro-positivos , temos

$$\mathbb{P}(VP) = \mathbb{P}(\text{classificado como } + | \text{é } +) = \frac{\mathbb{P}(\text{classif } + \text{ e é } +)}{\mathbb{P}(\text{é } +)}$$

- Esta probabilidade (estimada) é chamada de RECALL (revocação) em aprendizado de máquina ou de sensibilidade ou sensitividade em estudos epidemiológicos.
- Recall alto significa que o algoritmo retornou a maioria dos resultados relevantes.

Verdadeiro-negativos ou especificidade

- Quanto aos verdadeiro-negativos,

$$\mathbb{P}(VN) = \mathbb{P}(\text{classificado como -}|\text{é -}) = \frac{\mathbb{P}(\text{classif - e é -})}{\mathbb{P}(\text{é -})}$$

- Esta medida é chamada de *especificidade*.
- A idéia é que o algoritmo é específico para o que ele se propõe classficar.
- Se o item não é $+$, ele não retorna $+$.
- Veja que $\mathbb{P}(VN) + \mathbb{P}(FP) = 1$ pois um indivíduo que é negativo, será classificado ou como negativo (corretamente) ou como positivo (falsamente).
- Do mesmo modo, $\mathbb{P}(VP) + \mathbb{P}(FP) = 1$.

Estimando falso-positivos e falso-negativos

- Estimamos estas quantidades a partir dos dados comparando a verdadeira classe dos exemplos com a classe alocada a eles pela regressão logística.

	Diag -	Diag +
é -	429	71
é +	145	123

- Assim, o RECALL é estimado como

$$\mathbb{P}(VP) \approx \frac{123/768}{(145+123)/768} = \frac{123}{145+123} = 0.47$$

- Estamos acertando no diagnóstico de aprox metade dos verdadeiramente diabéticos.
- $\mathbb{P}(VN) \approx 429/(429+71) = 0.86$: acertamos mais frequentemente no diagnóstico dos verdadeiramente não-diabéticos.

Precisão, recall e especificidade

- Em aprendizado de máquina, uma métrica muito comum inverte os eventos usados na definição do RECALL.
- Temos RECALL igual a $\mathbb{P}(VP) = \mathbb{P}(\text{classificado como } + | \text{é } +)$.
- A PRECISÃO de um algoritmo de classificação é dada por

$$\text{Precisão} = \mathbb{P}(\text{é } + | \text{classificado como } +)$$

.

- Alta precisão indica que um algoritmo retornou mais resultados relevantes que irrelevantes.
- A partir da tabela anterior, podemos estimar a precisão como $123/(123 + 71) = 0.63$.
- Mais uma métrica, especificidade ($\mathbb{P}(VN) = \mathbb{P}(\text{classif -|e' -})$), estimada como $429/(429 + 71) = 0.86$.

Generalidade

- O método de máxima versossimilhança pode ser aplicado em praticamente toda situação de inferência em que os dados aleatórios sigam um modelo estatístico paramétrico $\mathcal{P}_{\boldsymbol{\theta}} = \{f(\boldsymbol{y};\boldsymbol{\theta})\}$.
- Isto é, os dados possuem uma distribuição de probabilidade que depende de um parâmetro desconhecido $\boldsymbol{\theta}$.
- Para modelos com um único parâmetro θ, o método pode ser resumido de maneira informal da seguinte maneira:
- Suponha que $y_1, \ldots, y_n$ são os dados da amostra.
- Usando o modelo estatístico $\mathcal{P}_{\theta}$, calcule o valor aproximado da probabilidade de observar os dados da amostra e obtenha a *função de verossimilhança* $L(\theta)$ onde apenas θ pode variar.
- Obtenha o valor $\hat{\theta}$ que maximiza $L(\theta)$. Este valor é a estimativa de máxima verossimilhança.

Por quê usar o método de máxima verossimilhança?

- generalidade: o método é muito geral e pode ser usado quando a intuição não conseguir sugerir bons estimadores para θ.
- É fácil obter $L(\theta)$ e basta maximizá-la em θ.
- **Fisher:** se a amostra cresce então a estimativa $\hat{\theta}$ converge para θ QUALQUER QUE SEJA O PROBLEMA ESTATÍSTICO.
- **Fisher:** se a amostra cresce então a estimativa $\hat{\theta}$ é aproximadamente não-viciada para θ.
- OBS: Um estimador é não-viciado se as estimativas que fazemos com ele tendem a oscilar em torno do verdadeiro valor desconhecido de θ (veremos isto mais a frente).

Por quê usar o método de máxima verossimilhança?

- outra razão para usar a estimativa de máxima verossimilhança.
- Esta razão també é de **Fisher**, e o resultado é sensacional: <u>qualquer</u> estimador não-viciado ou aproximadamente não-viciado terá um erro médio de estimação maior que o estimador de máxima verossimilhança. E isto é válido para praticamente *qualquer* modelo estatístico.
- **Fisher** de novo: o estimador de máxima verossimilhança possui distribuição aproximadamente normal, não importa quão complicada seja a sua fórmula. Este é um fato fundamental para intervalos de confiança e testes de hipóteses.