

Aprendizagem de Máquina 1

Inteligência Artificial



AULA 03 — REGRESSÃO LOGÍSTICA

Larissa Driemeier
Thiago Martins

CRONOGRAMA

Data	Professor	Assunto
07/05	Larissa	Definição de aprendizado de máquina. Aprendizado supervisionado e não supervisionado. Regressão linear. Regressão polinomial.
14/05	Thiago	Exercícios de acompanhamento. Nota 01.
21/05	Larissa	Regressão Logística
28/05	Thiago	Exercícios de acompanhamento. Nota 02.
04/06	Larissa	Máquinas de vetores de suporte
11/06	Thiago	Exercícios de acompanhamento. Nota 03.
18/06	Larissa	Aprendizado não supervisionado
25/06	Thiago	Exercícios de acompanhamento. Nota 04.
02/07	Larissa	Redução de similaridade: análise de componentes principais (PCA) e suas variações.
16/07	Larissa/Thiago	Exercícios “Melhores Momentos”. Nota 05.

AULA DE HOJE

- Revisão da última aula
- Regressão logística
 - O que é isso? Porque é chamada de *logística*?
 - Onde aplicar?
 - Classificação binária vs multi classe



REVISÃO

Modelos lineares de regressão

MODELO DE REGRESSÃO

HIPÓTESE:

$$h_{\omega}(x) = \hat{y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 = \omega^T x$$

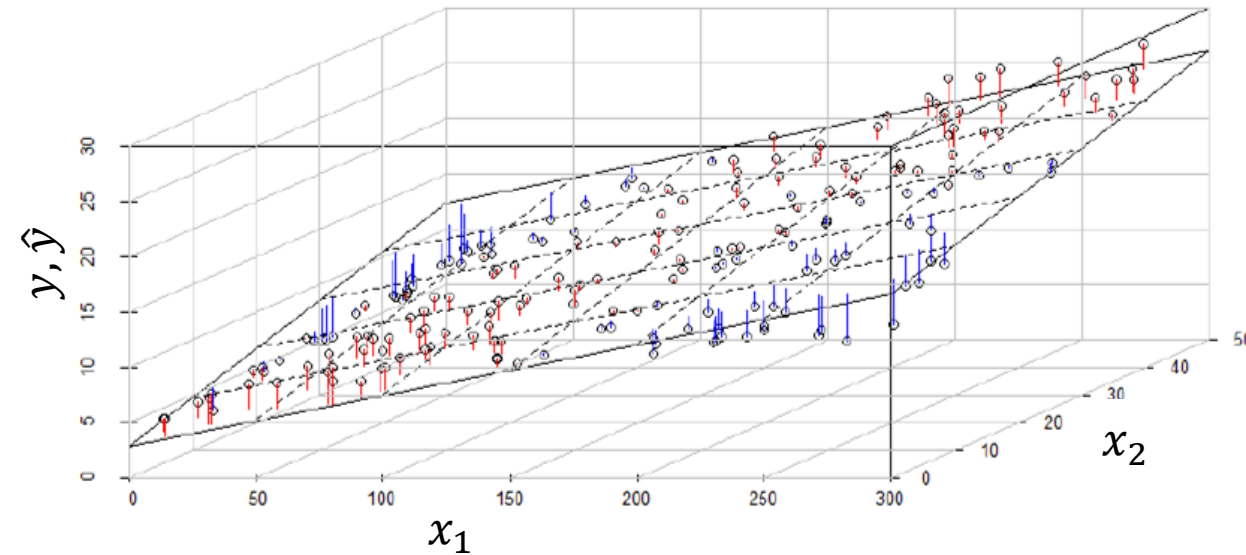
FUNÇÃO CUSTO:

$$J(\omega_0, \omega_1, \omega_2) = \frac{1}{2m} \sum_{i=1}^m [\epsilon^{(i)}]^2, \epsilon^{(i)} = \hat{y}^{(i)} - y^{(i)}$$

GOL:

$$\min_{(\omega_0, \omega_1, \omega_2)} J(\omega_0, \omega_1, \omega_2)$$

ONE STEP LEARNING: $\omega = (X^T X)^{-1} X^T y$



Transformação não linear

$\omega^T x$ é linear em ω

Qualquer transformação $x \xrightarrow{\Phi} z$ preserva esta linearidade

Exemplo: $(x_1, x_2) \xrightarrow{\Phi} (x_1^2, x_2^2)$



REGRESSÃO LOGÍSTICA

O que é isso?

REGRESSÃO VS CLASSIFICAÇÃO

Um problema de regressão tem um número real como saída.

Por exemplo, podemos usar os dados da tabela para estimar o peso de alguém de acordo com sua altura.

Altura (cm)	Peso(kg)
167,1	51,3
181,7	61,9
176,3	69,4
173,3	64,6
172,2	65,5
174,5	55,9
177,3	64,2
177,8	61,9
172,5	51,0
169,6	54,7
168,9	57,8
171,8	51,8
173,5	57,0
170,5	55,5
173,4	52,7

Um problema de classificação tem um valor discreto como saída.

Por exemplo, “gosta de abacaxi na pizza” e “não gosta de abacaxi na pizza” são opções discretas. Não há meio termo. Este é um exemplo de classificação binária.

A regressão logística é usada quando a variável dependente (alvo) é categórica.

Idade	Gosta de abacaxi na pizza
42	1
65	1
50	1
76	1
96	1
50	1
91	0
58	1
25	1
23	1
75	1
46	0
87	0
96	0
45	0
32	1
63	0
21	1
26	1
93	0
68	1
96	0

REGRESSÃO LOGÍSTICA

Hoje apresentaremos um algoritmo importante que é admiravelmente adequado para descobrir a ligação entre características ou pistas e algum resultado específico: regressão logística.

A regressão logística é o algoritmo básico de aprendizado de máquina supervisionado para classificação e também tem uma relação muito estreita com redes neurais. Na verdade, uma rede neural pode ser vista como uma série de classificadores de regressão logística empilhados uns sobre os outros.

A regressão logística pode ser usada para classificar uma observação em uma de duas classes (como 'sentimento positivo' e 'sentimento negativo') ou em uma de muitas classes.

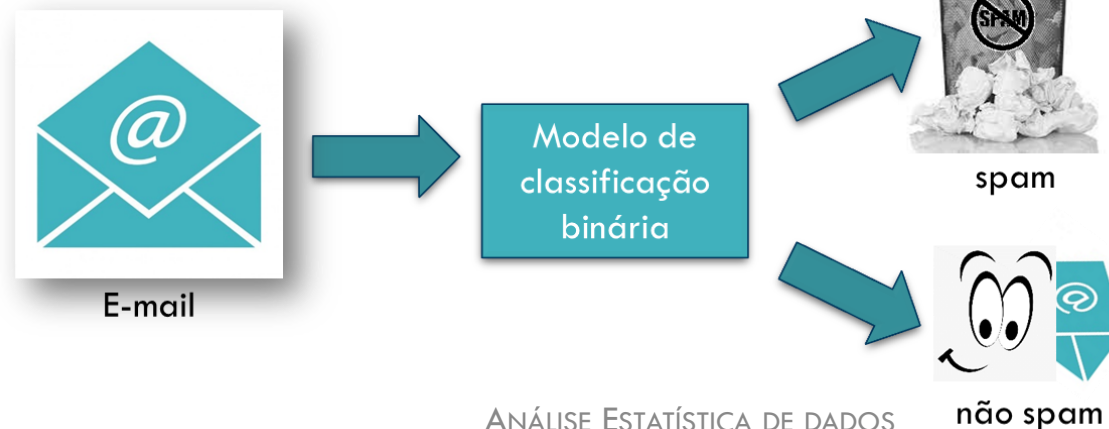
"E como você sabe que essas lindas begônias não têm a mesma importância?"

Hercule Poirot, em O Misterioso Caso de Styles, de Agatha Christie

REGRESSÃO LOGÍSTICA BINÁRIA

Como a matemática para o caso de duas classes é mais simples, descreveremos esse caso especial de regressão logística primeiro nas próximas seções e, em seguida, resumiremos brevemente o uso da regressão logística multinomial para mais de duas classes.

O modelo binário fornece um resultado dicotômico limitado a duas possíveis saídas: sim/não, 0/1 ou verdadeiro/falso.



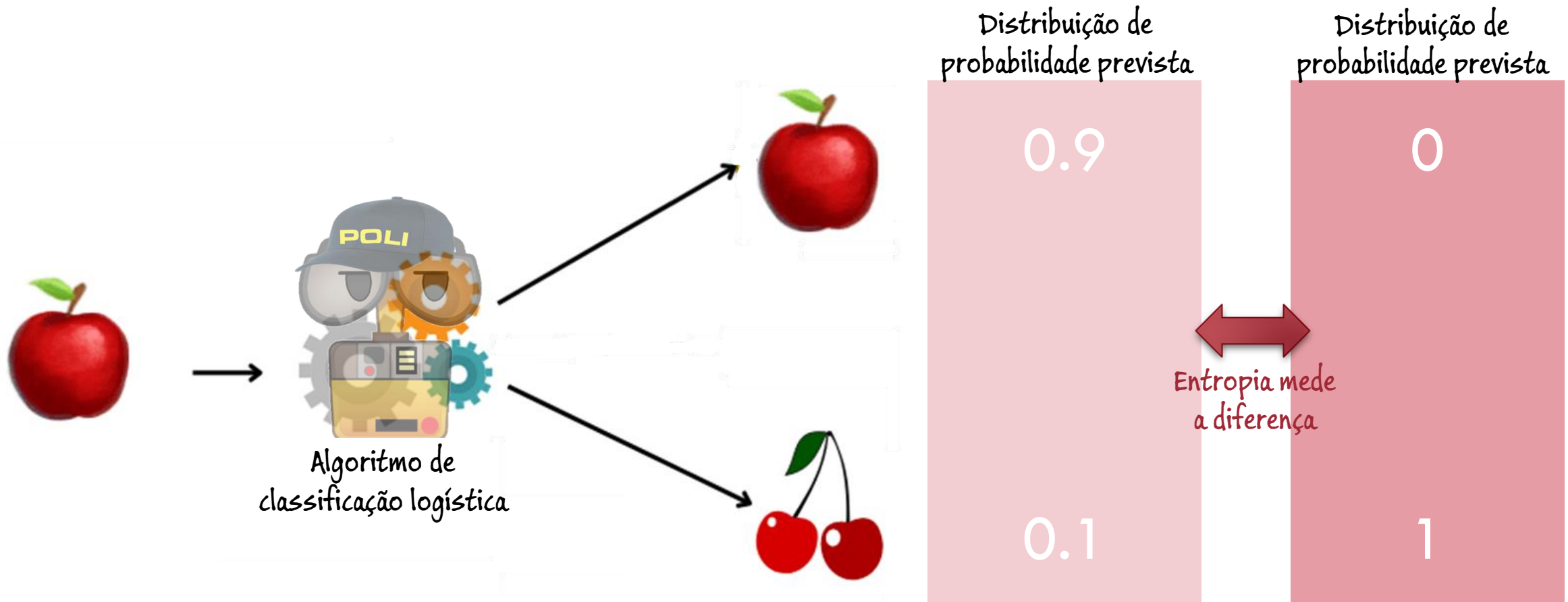
Queremos prever se um e-mail é spam(1) ou não(0). Isto é, qual a probabilidade de ser um spam, e, a partir de uma probabilidade limite, decidimos se este será classificado como spam ou não.

TÉCNICA DE REGRESSÃO LOGÍSTICA

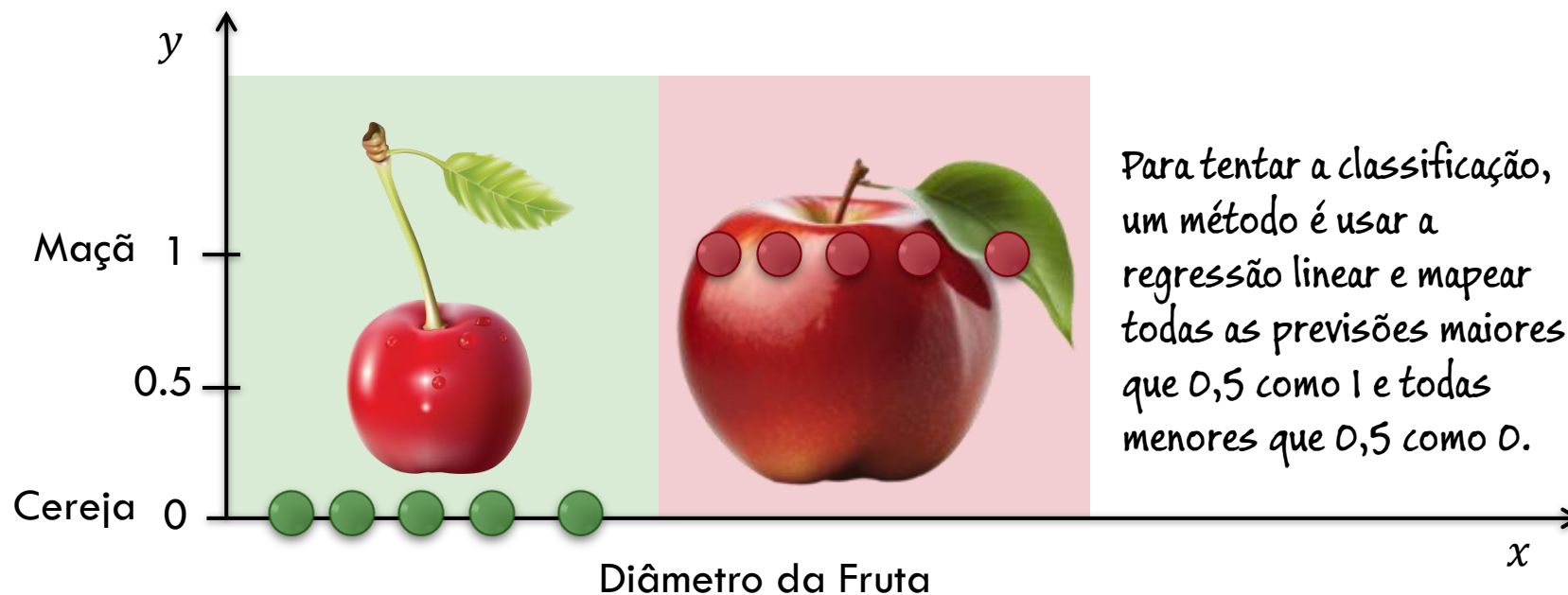
Nessa linha, temos outros exemplos:

- queremos checar exames radiológicos para verificar se um tumor é maligno (1) ou não (0);
- queremos saber se o passo de seu exoesqueleto é estável (1) ou não (0), a partir dos movimentos das juntas;
- queremos analisar os registros bancários históricos para prever se um cliente deixará de pagar seus empréstimos (1) ou pagará o empréstimo (0).
- Para negócios digitais em particular, o algoritmo pode ser usado para recomendações de produtos, classificação de leads, antecipação de eventos raros como inadimplência...

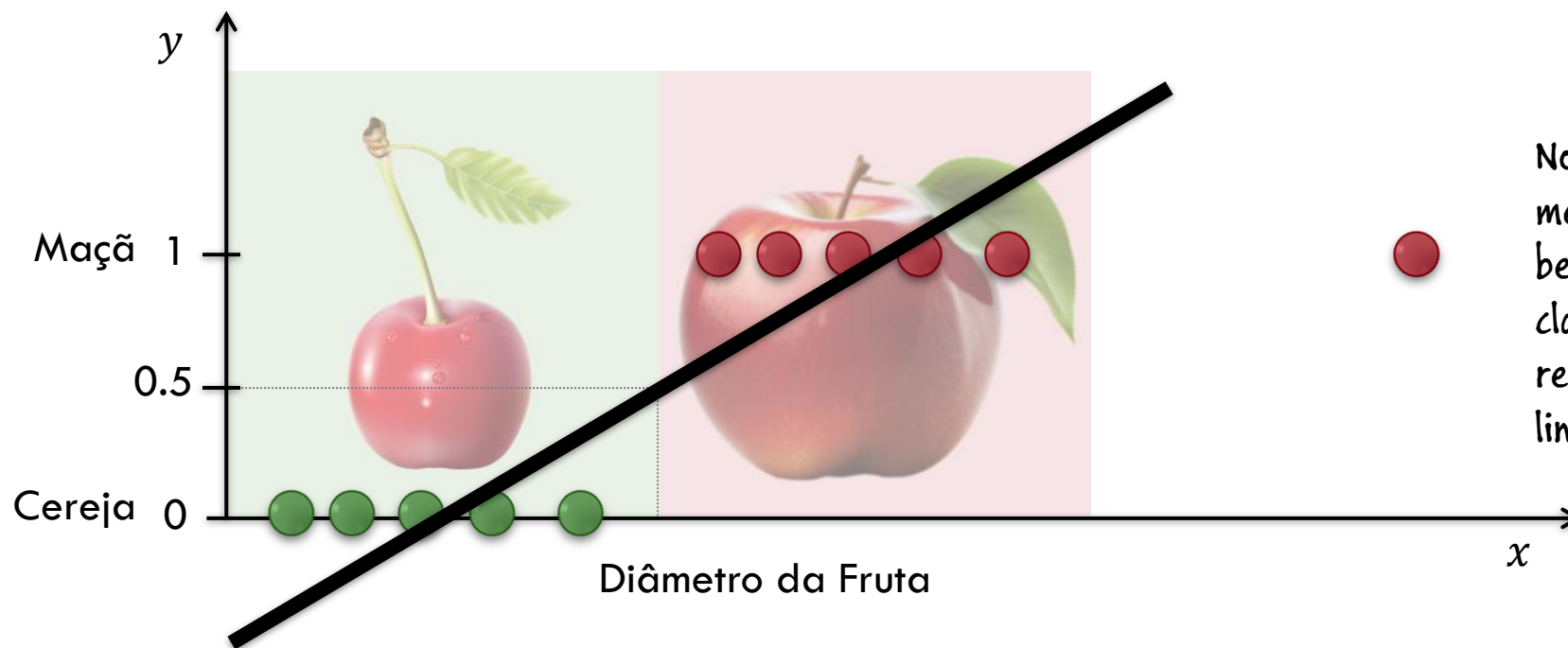
VAMOS DIFERENCIAR CEREJA DE MAÇÃ



MAÇÃ OU CEREJA?

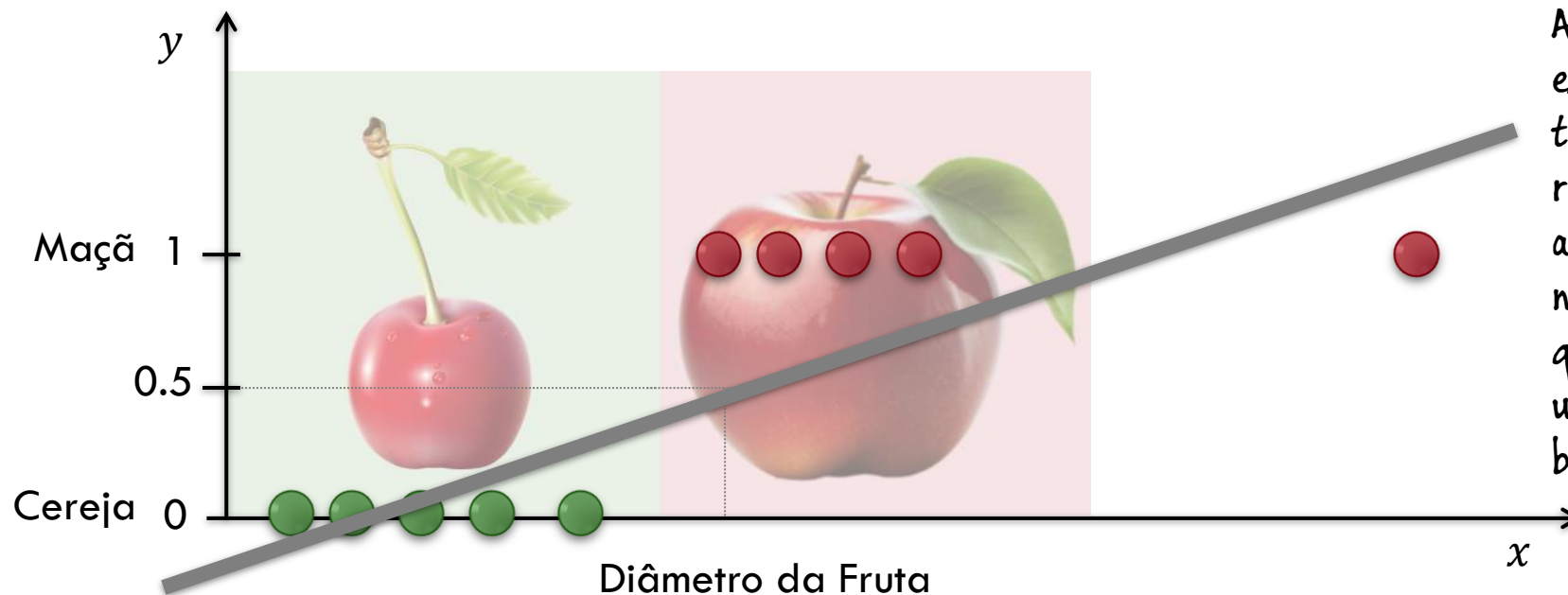


MAÇÃ OU CEREJA?



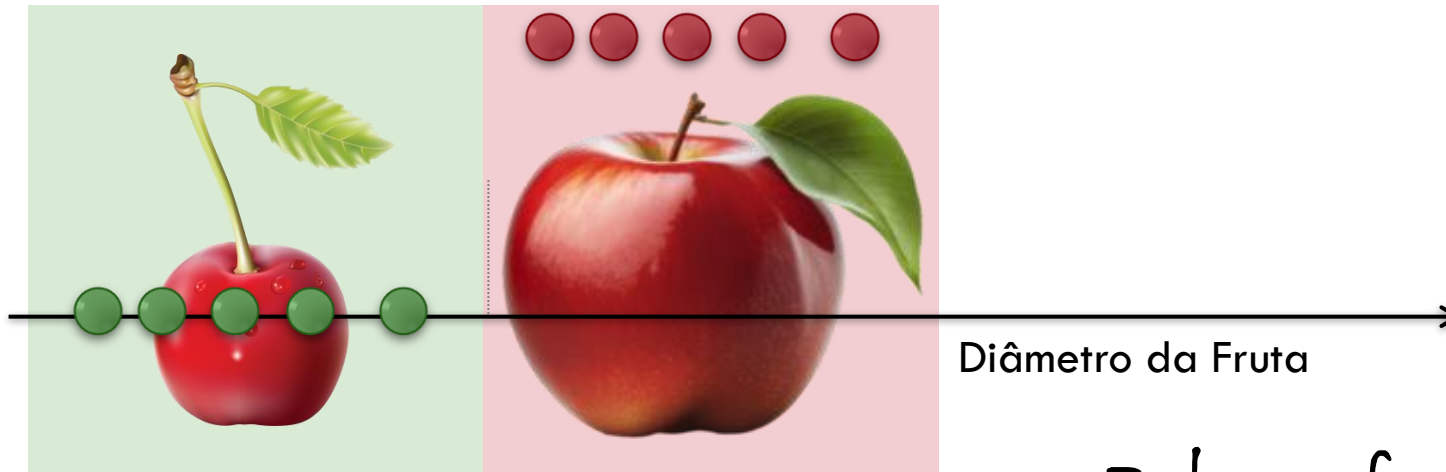
No entanto, esse método não funciona bem porque a classificação não é realmente uma função linear.

MAÇÃ OU CEREJA?



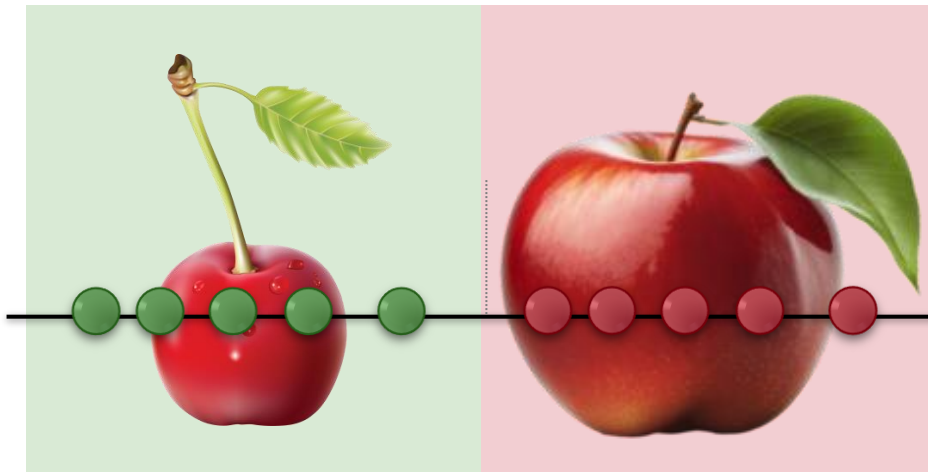
Antes de adicionar este exemplo extra de treinamento, a regressão linear estava apenas tendo sorte e nos deu uma hipótese que funcionou bem para um conjunto de dados bem específico.

MAÇÃ OU CEREJA?



Podemos formular o problema da seguinte forma: “a fruta é maçã?” ou, melhor ainda, “qual a probabilidade da fruta ser maçã?”

MAÇÃ OU CEREJA?

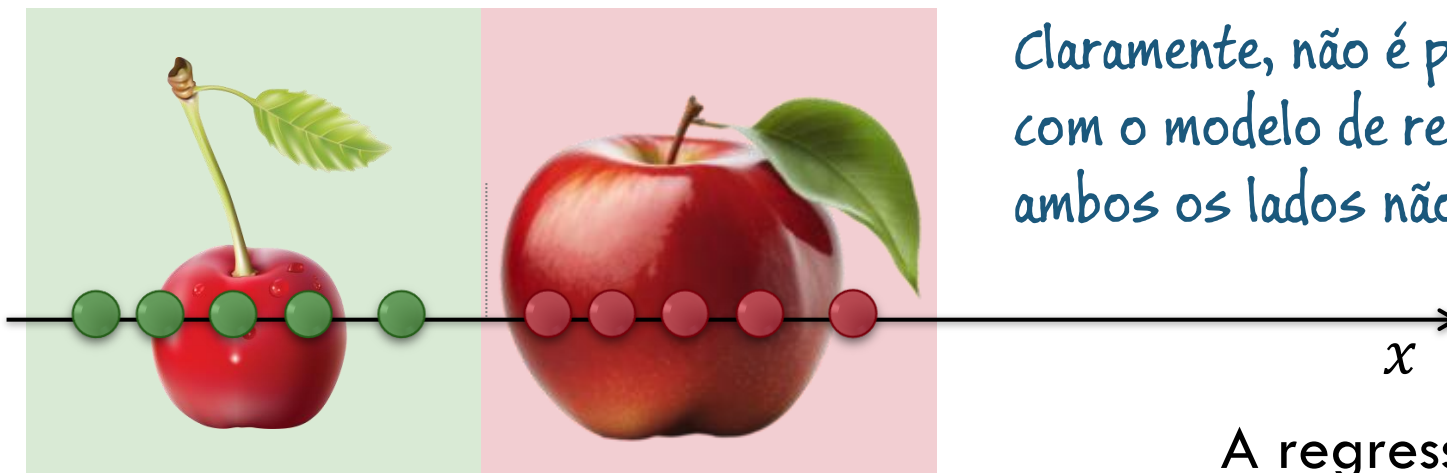


Teoricamente, **maças** deveriam ter uma probabilidade de **1.0** (de serem maçãs), ao passo que **cerejas** deveriam ter uma probabilidade de **0.0** (de serem maçãs).

Diâmetro da Fruta

Assim, **maças** pertencem à **classe positiva (SIM, 1, elas são maçãs)**, e **cerejas** pertencem à classe negativa (**NÃO, 0, elas não são maçãs**).

MAÇÃ OU CEREJA?



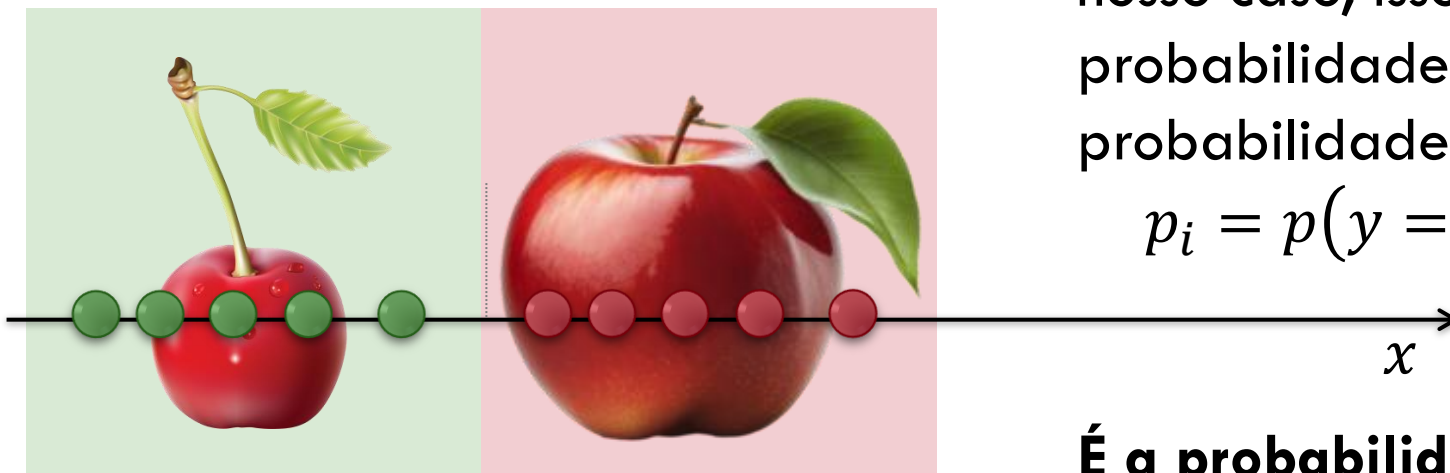
Claramente, não é possível obter esse resultado com o modelo de regressão, pois o intervalo de ambos os lados não corresponde.

$$h(x) = \hat{y} = \omega^T x$$

$$\hat{y}_i = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \dots + \omega_n x_n^{(i)}$$

A regressão logística é sobre classificação, ou seja, \hat{y}_i é uma variável categórica – valores 0,1. O lado direito é governado pelo intervalo de variáveis de recurso, $[-\infty, +\infty]$.

MAÇÃ OU CEREJA?



O resultado, \hat{y}_i , assume o valor 1 (em nosso caso, isso representa maçã) com probabilidade p_i e o valor 0 com probabilidade $1 - p_i$

$$p_i = p(y = k | x^{(i)}; \omega), \quad k = 0, 1$$

É a probabilidade p_i que modelamos em relação às variáveis independentes.

$$h(x) = \hat{y} = \omega^T x$$

$$\boxed{\hat{y}_i \rightarrow p_i} = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \dots + \omega_n x_n^{(i)}$$

transformação

Queremos escolher uma transformação na equação acima que faça sentido prático e matemático.

RAZÃO DE CHANCES (ODDS RATIO)

A razão de chances descreve a relação entre a probabilidade de sucesso e de falha. Se um evento ocorre com probabilidade p , a razão de chances deste evento é

$$OR = \frac{\text{probabilidade sim}}{\text{probabilidade não}} = \frac{p}{1 - p}$$

Exemplo: Razão de chances de se obter “2” ao jogar um dado honesto,

$$OR = \frac{\text{probabilidade sim}}{\text{probabilidade não}} = \frac{1/6}{5/6} = \frac{1}{5} = 1:5$$

VEJAM ESTE OUTRO EXEMPLO...

	Não compra	Compra	Total
Mulher	106	159	265
Homem	125	120	245

$$OR_M = \frac{159/265}{106/265} = 1.5 = 3:2$$

$$OR_H = \frac{120/245}{125/245} = 0.96 = 24:25$$

OR pode variar de 0 a ∞ . Quanto maior OR, melhor é a chance de sucesso.

A primeira transformação seria dividir p_i por $1 - p_i$,

$$OR_i = \frac{p_i}{1 - p_i} = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \dots + \omega_n x_n^{(i)}$$

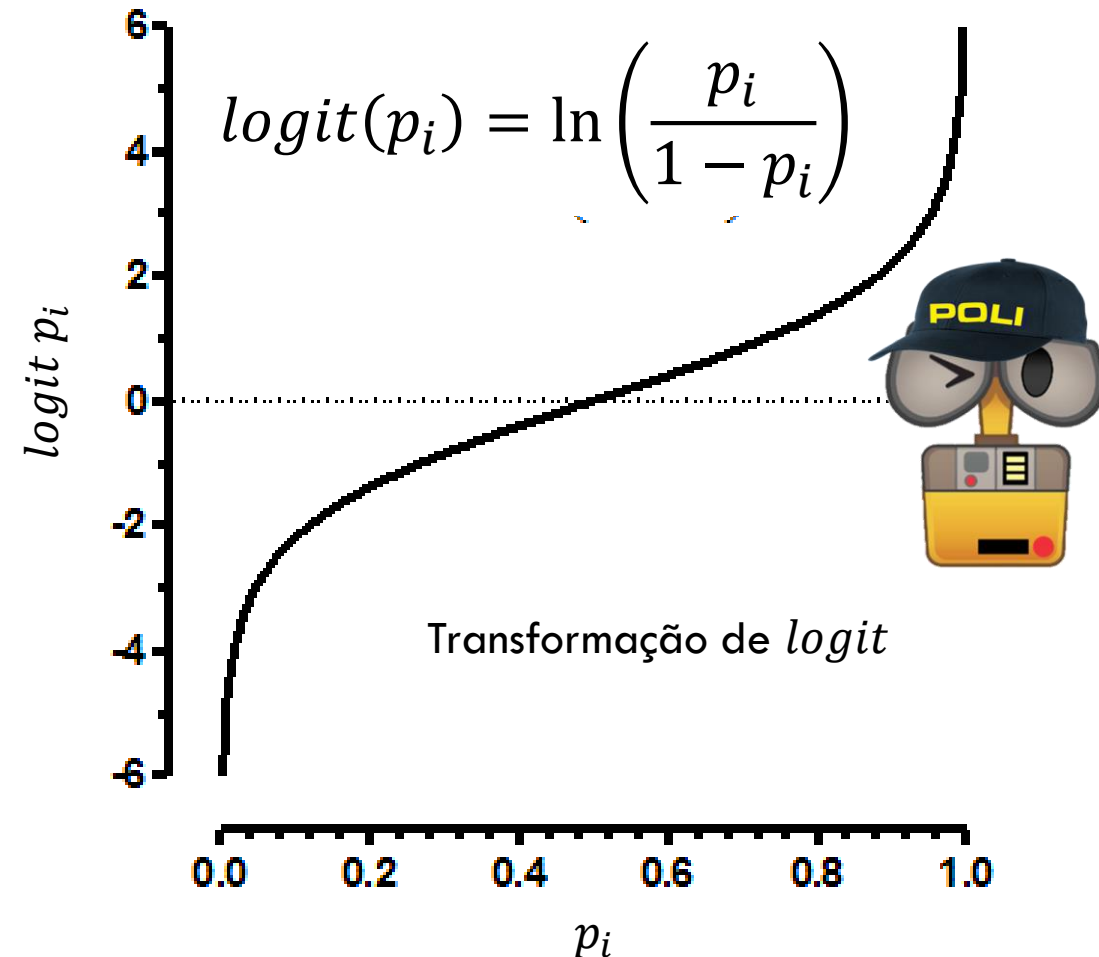


TRANSFORMAÇÃO LOGIT

Na segunda transformação, se aplicarmos a função \ln a $p_i/1 - p_i$, então encontramos uma transformação que torna o intervalo de possibilidades do lado esquerdo da equação igual ao intervalo de possibilidades do lado direito,

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \omega^T x$$

Sabendo que \ln de um valor que tende a 0, tende a $-\infty$ e \ln de um valor que tende a infinito tende a $+\infty$.



O PROBLEMA DE CLASSIFICAÇÃO

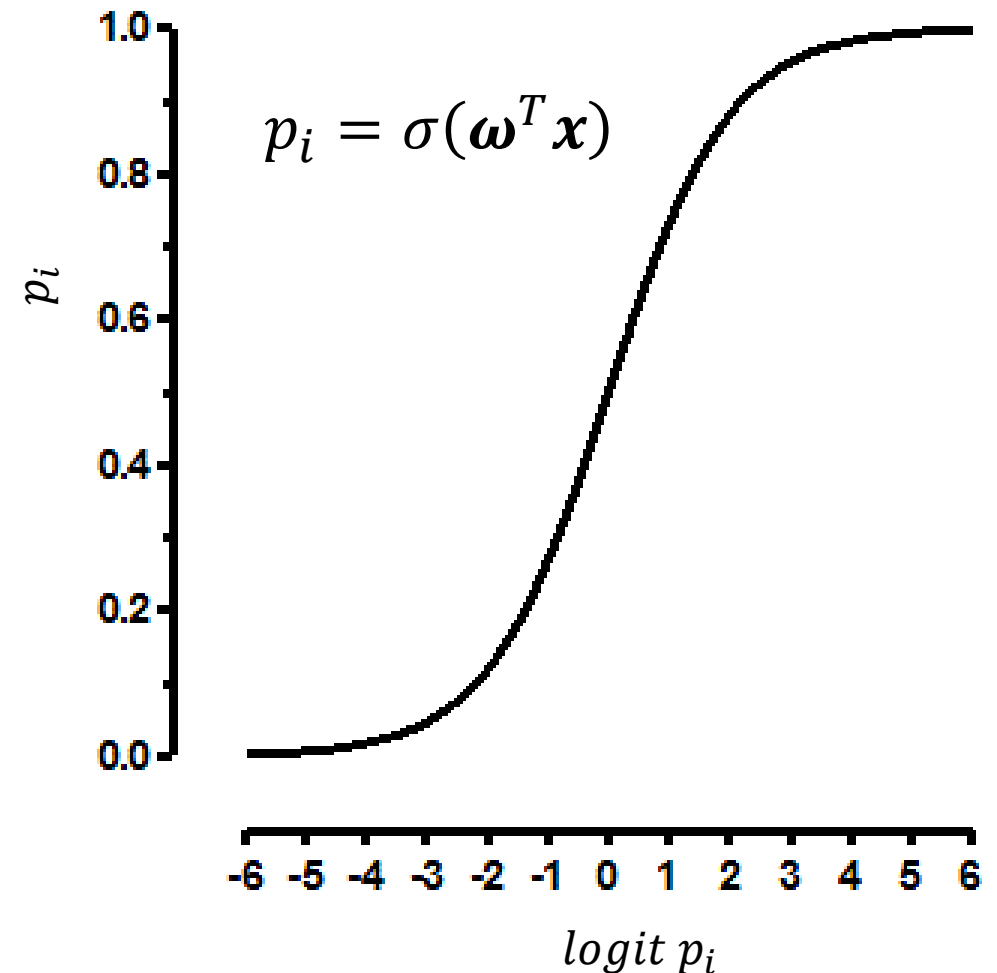
$$\ln \left(\frac{p_i}{1 - p_i} \right) = \omega^T x$$

Resolvendo para p_i tal que,
$$p_i = \sigma(\omega^T x)$$

$0 \leq \sigma(\omega^T x) \leq 1$ para qualquer x .
é um modelo de probabilidade

$$p(y = k | x^{(i)}; \omega) = \sigma(\omega^T x^{(i)})$$

descreve a probabilidade para o dado i pertencer à classe $k = 1, 2$, dado que conhecemos a entrada x , parametrizada por ω .

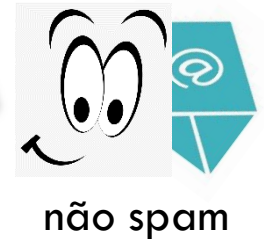


CLASSIFICAÇÃO BINÁRIA

Como falamos, por enquanto, vamos nos concentrar no problema de classificação binária no qual y pode assumir apenas dois valores, 0 e 1.



Modelo de
classificação
binária



Para problemas de classificação binária ($k = 1, 2$, e $y = 0, 1$), aprenderemos um modelo $\sigma(x)$ para o qual

$$p(y = 1|x; \omega) \text{ é modelado por } \sigma(\omega^T x)$$
$$p(y = 0|x; \omega) \text{ é modelado por } 1 - \sigma(\omega^T x)$$

RESUMO: REGRESSÃO LOGÍSTICA BINÁRIA

$$y^{(i)} = 0,1$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \dots + \omega_n x_n^{(i)}$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \omega^T x$$

$$e^{\ln\left(\frac{p_i}{1-p_i}\right)} = e^{\omega^T x}$$

$$\frac{p_i}{1-p_i} = e^{\omega^T x}$$

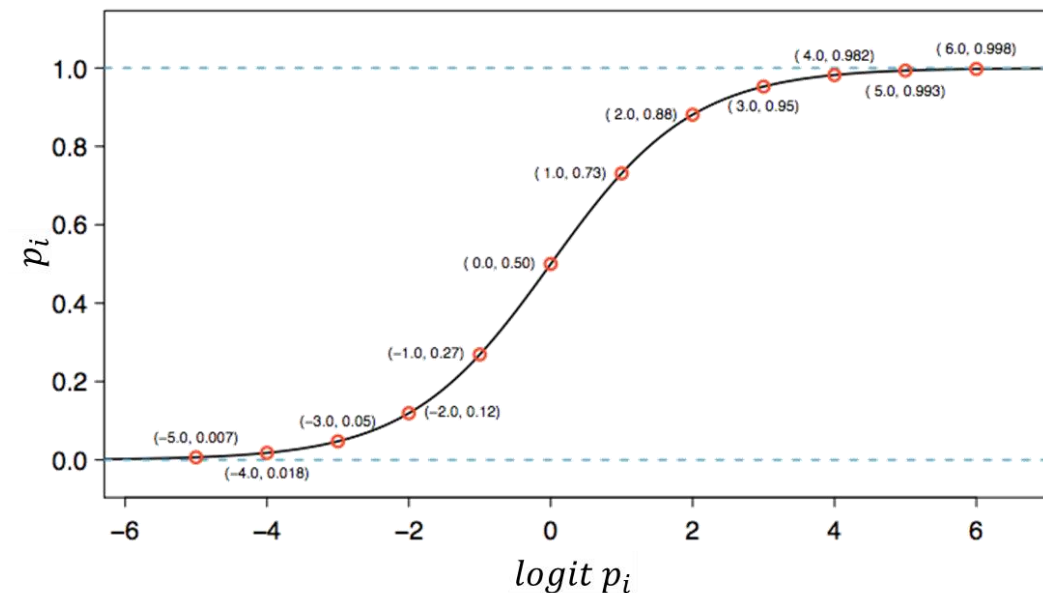
$$p_i = (1-p_i)e^{\omega^T x}$$

$$p_i + p_i e^{\omega^T x} = e^{\omega^T x}$$

$$p_i (1 + e^{\omega^T x}) = e^{\omega^T x}$$

$$p_i = \frac{e^{\omega^T x}}{1 + e^{\omega^T x}} \frac{e^{-\omega^T x}}{e^{-\omega^T x}}$$

$$p_i = \frac{1}{1 + e^{-\omega^T x}}$$



RESUMO: REGRESSÃO LOGÍSTICA BINÁRIA

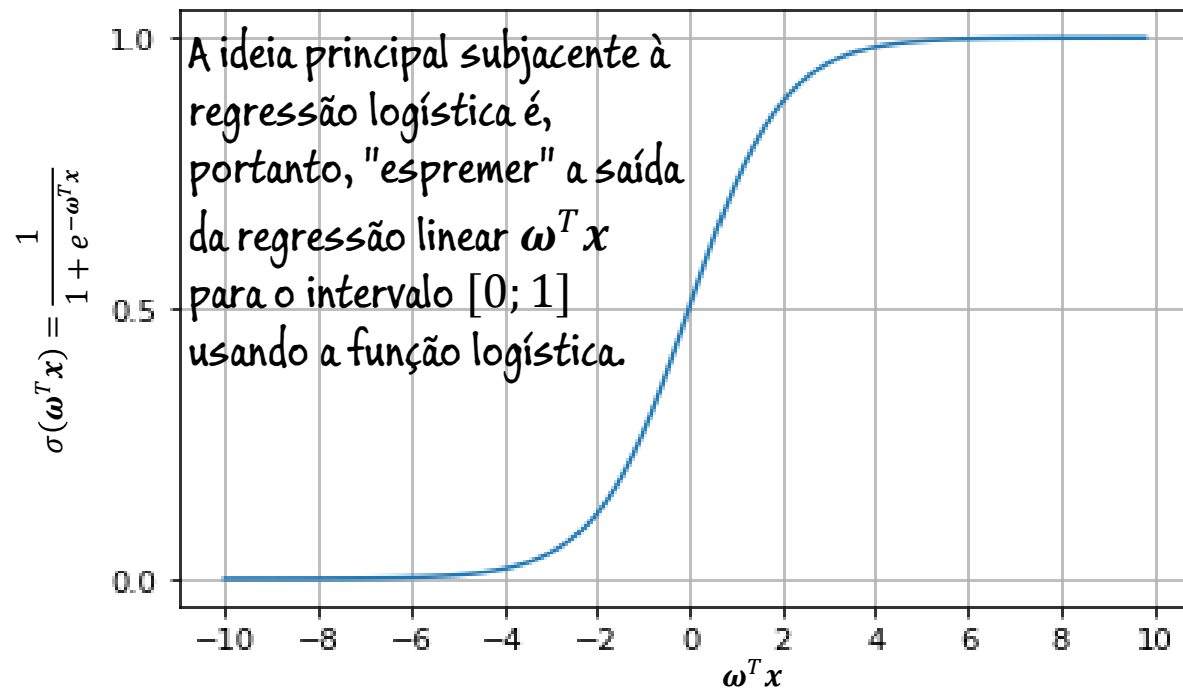
$$y^{(i)} = 0,1$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \dots + \omega_n x_n^{(i)}$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \omega^T x$$

$$p_i = \sigma(\omega^T x) = \frac{1}{1 + e^{-\omega^T x}}$$

Temos uma hipótese de como modelar nosso problema.



$$h(x) = \sigma(\omega^T x) = \frac{1}{1 + e^{-\omega^T x}}$$

A curva $s(x) = \frac{1}{1+e^{-x}}$ é também conhecida como sigmoide, que é a inversa da logit.

RESUMO: REGRESSÃO LOGÍSTICA BINÁRIA

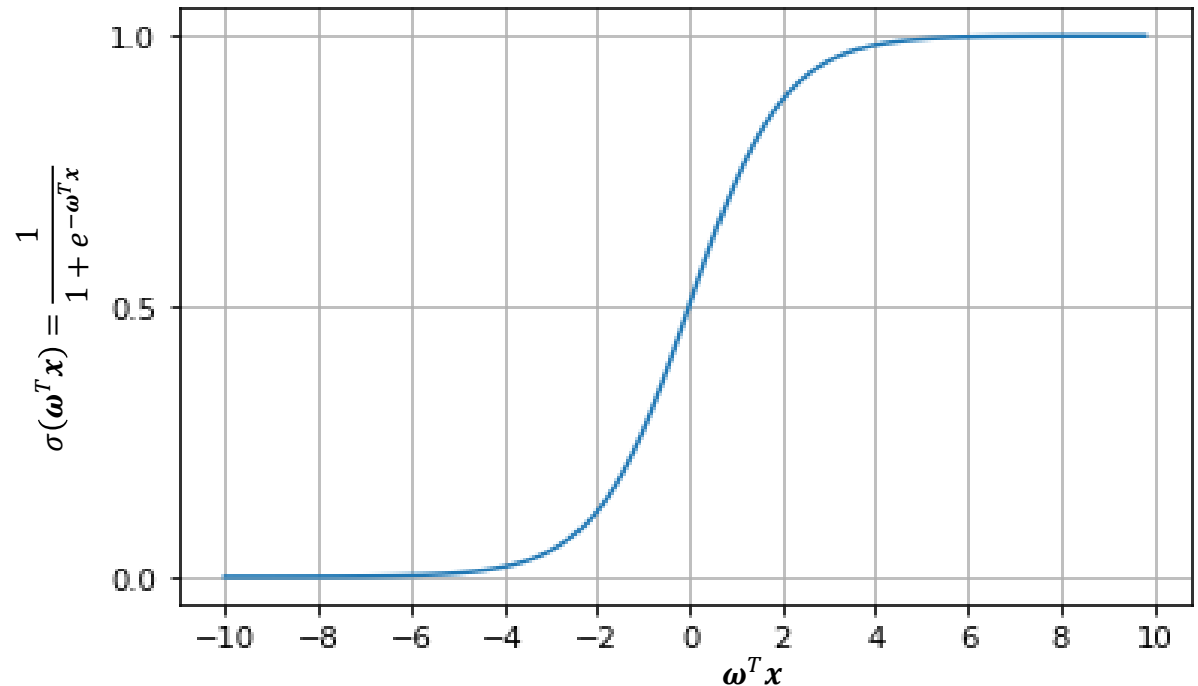
$$y^{(i)} = 0,1$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \dots + \omega_n x_n^{(i)}$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\omega}^T \mathbf{x}$$

$$p_i = \sigma(\boldsymbol{\omega}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\omega}^T \mathbf{x}}}$$

Temos uma hipótese de como modelar nosso problema.



$$h(\mathbf{x}) = \sigma(\boldsymbol{\omega}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\omega}^T \mathbf{x}}}$$

Agora temos um modelo que contém parâmetros desconhecidos. Como na regressão linear, os parâmetros podem ser aprendidos a partir de dados de treinamento..

VAMOS POR PARTES...

Primeiro vamos
interpretar o
que significa
matematicamente



$$h(x) = \sigma(\omega^T x) = \frac{1}{1 + e^{-\omega^T x}}$$

Depois vamos
entender como
encontrar
 ω





A HIPÓTESE h

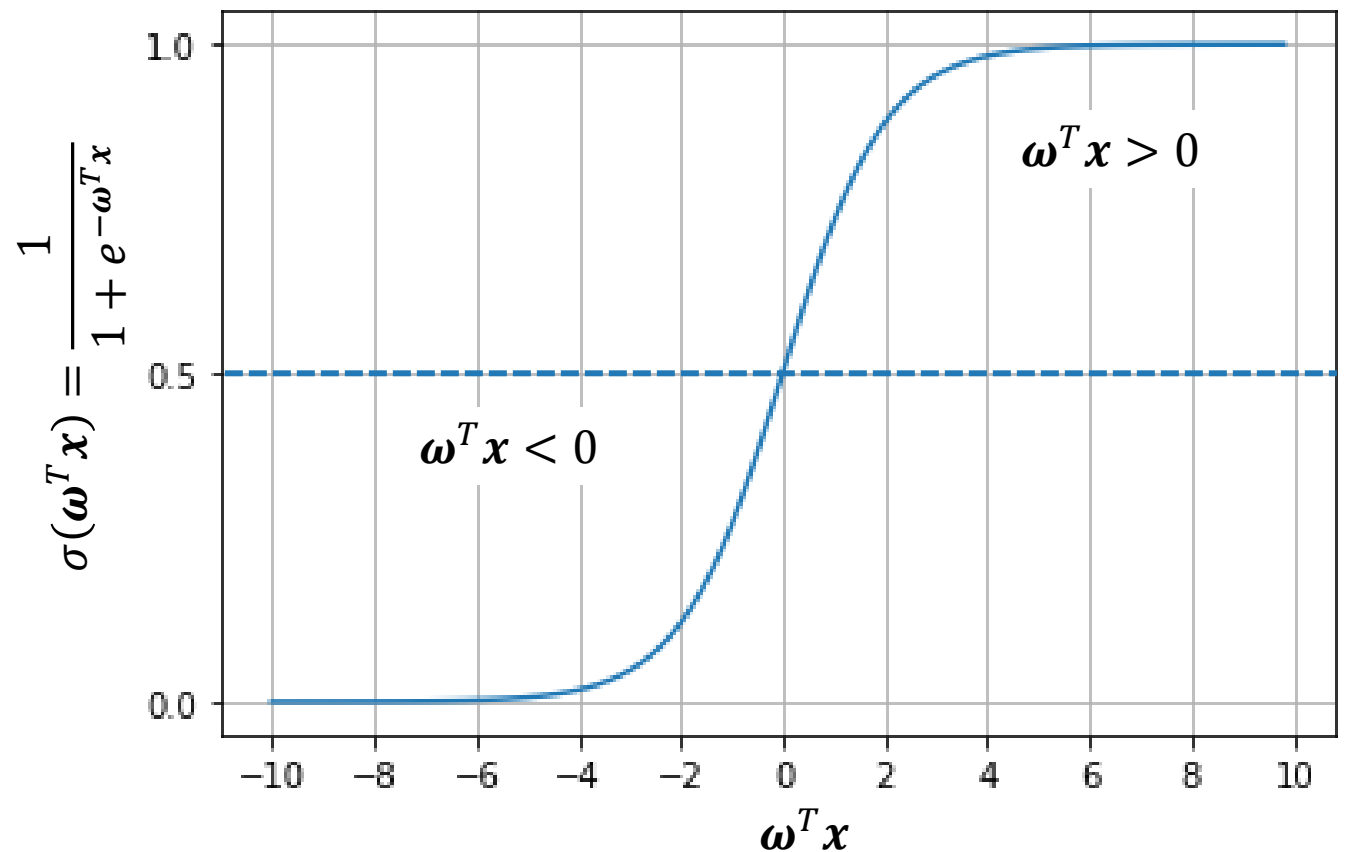
$$h(x) = \sigma(\omega^T x) = \frac{1}{1 + e^{-\omega^T x}}$$

HIPÓTESE h

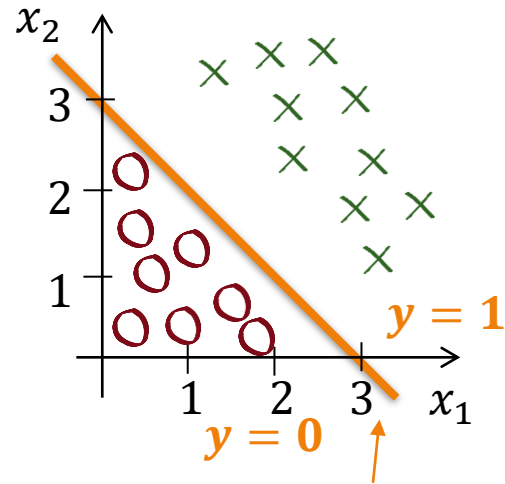
$$h_{\omega}(\mathbf{x}) = \frac{1}{1 + e^{-\omega^T \mathbf{x}}} = \sigma(\omega^T \mathbf{x})$$

$$h_{\omega}(\mathbf{x}) = \begin{cases} > 0.5 & \omega^T \mathbf{x} > 0 \\ < 0.5 & \omega^T \mathbf{x} < 0 \end{cases}$$

Se a soma ponderada de entradas for maior que zero, a classe prevista é 1, caso contrário, será 0. Portanto, **o limite de decisão que separa as duas classes pode ser encontrado configurando a soma ponderada das entradas como $\omega^T \mathbf{x} = 0$.**



CONTORNO DE DECISÃO



Contorno de decisão

É uma propriedade da hipótese e dos parâmetros definidos.

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \end{bmatrix} = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

$$h_{\boldsymbol{\omega}}(\mathbf{x}) = \sigma(\boldsymbol{\omega}^T \mathbf{x}) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2)$$

Prevemos $y = 1$ se:

$$-3 + x_1 + x_2 \geq 0$$

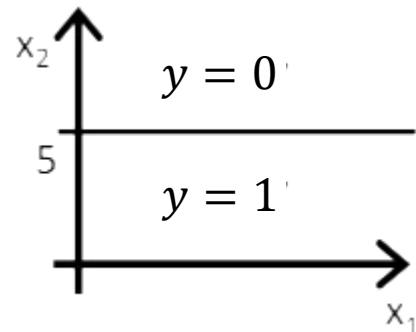
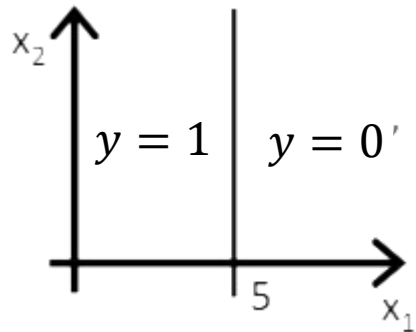
$$x_1 + x_2 = 3 \quad h_{\boldsymbol{\omega}}(\mathbf{x}) = 0.5$$

QUESTÃO PARA VOCÊ PENSAR

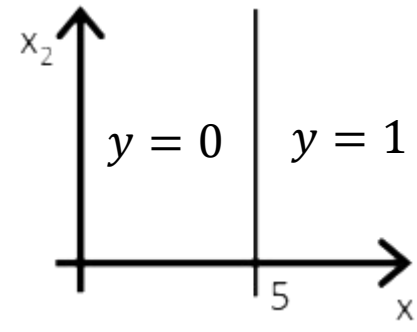
Considere a regressão logística com 2 características,

$$h_{\omega}(\mathbf{x}) = \sigma(\boldsymbol{\omega}^T \mathbf{x}) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2).$$

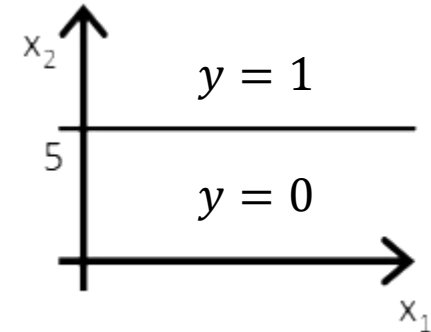
Suponha $\omega_0 = 5, \omega_1 = -1, \omega_2 = 0$, isto é, $h_{\omega}(\mathbf{x}) = \sigma(5 - x_1)$.



(b)

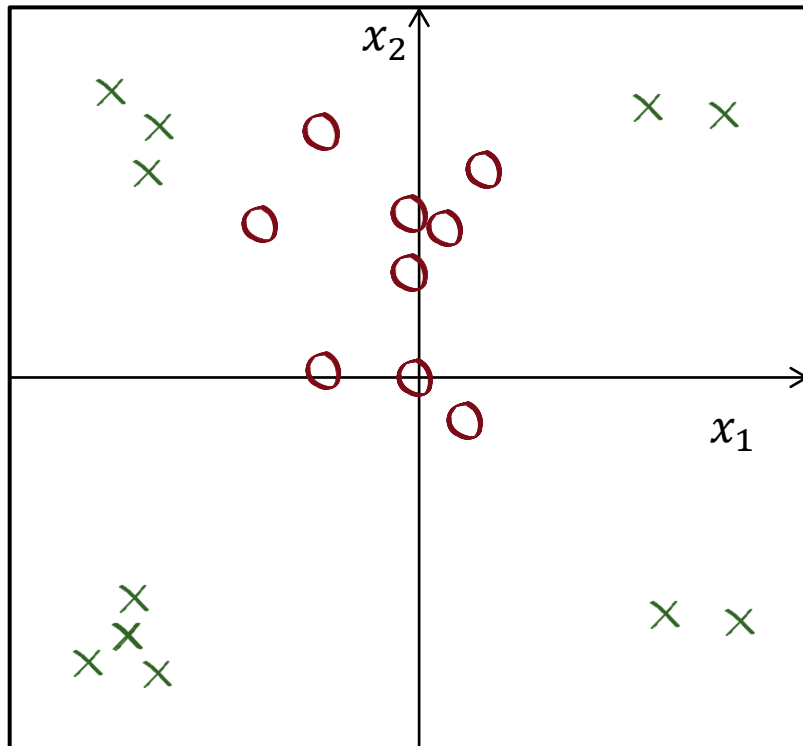


(c)



(d)

LINEARIDADE É LIMITADA!



Tente colocar a reta em qualquer lugar... Sempre será um problema, porque os dados não são separáveis linearmente!

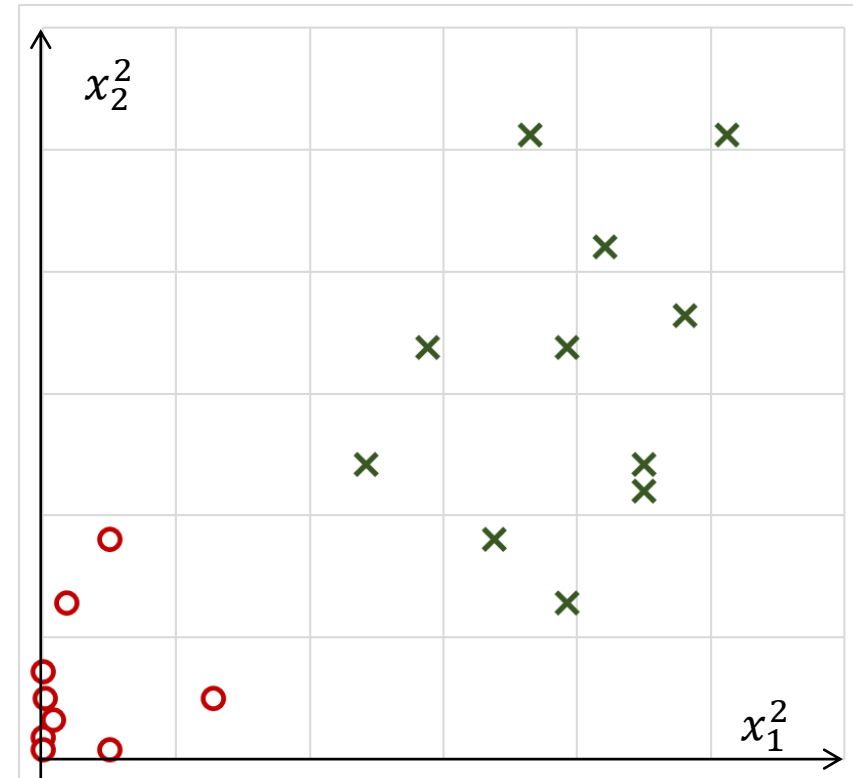
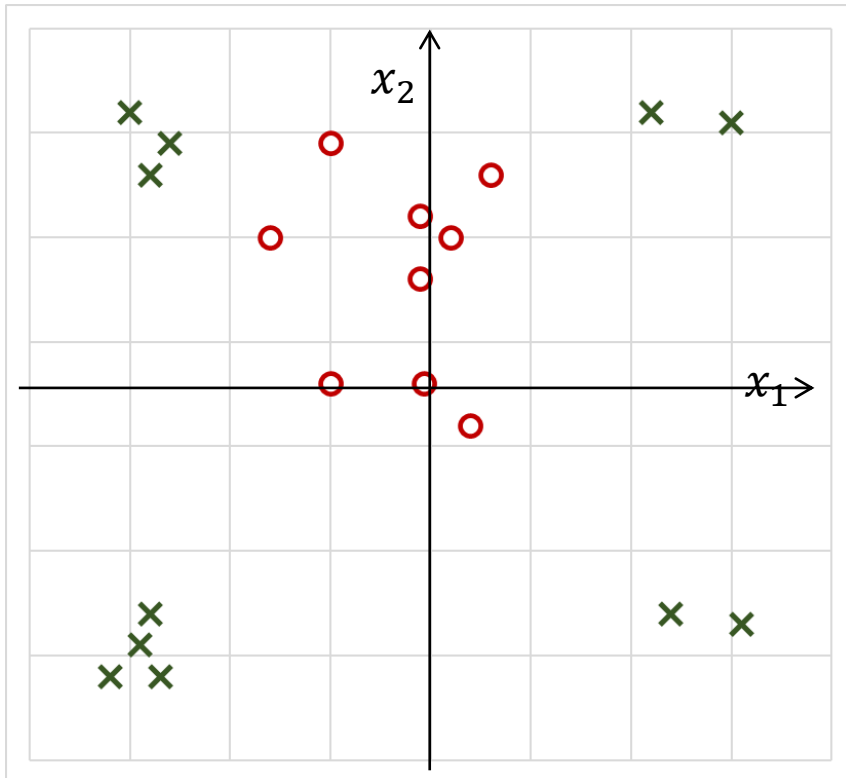
Lembram-se da aula anterior???

Na expressão polinomial $\omega^T x$, nosso problema deve ser linear com relação aos pesos. Os dados são constantes... **E isso abriu um universo!** Podemos fazer transformações não lineares incríveis com os dados. Eles viram dados mais elaborados, mas constantes!!!

Os pesos que daremos às características não lineares têm ainda uma dependência linear.

TRANSFORMAR...

$$(x_1, x_2) \xrightarrow{\Phi} (x_1^2, x_2^2)$$



CONTORNO NÃO LINEAR

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

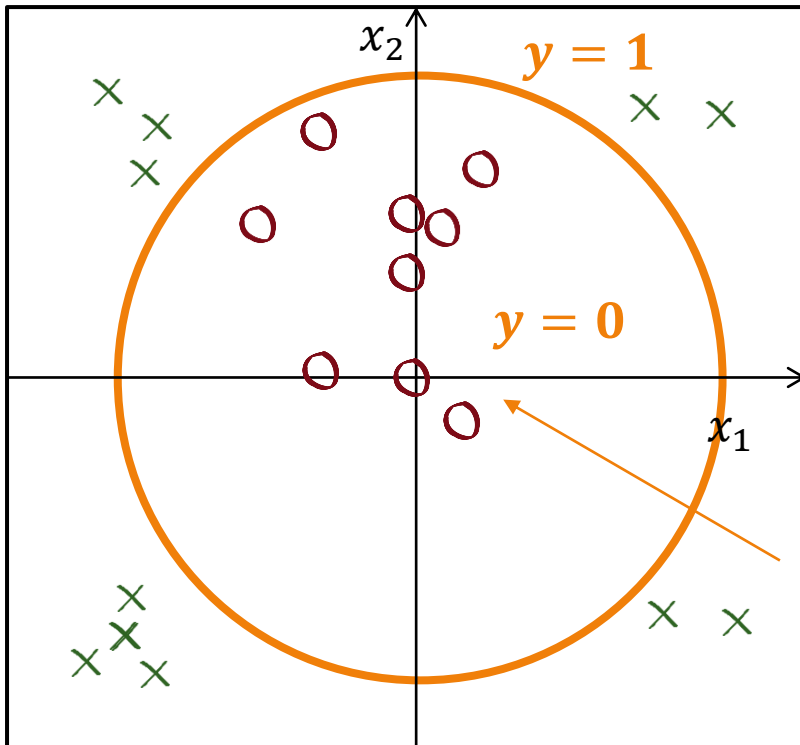
$$h_{\boldsymbol{\omega}}(\mathbf{x}) = \sigma(\boldsymbol{\omega}^T \mathbf{x}) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2)$$

Preveremos $y = 1$ se:

$$-1 + x_1^2 + x_2^2 \geq 0$$

Contorno de decisão

$$x_1^2 + x_2^2 = 1 \quad h_{\boldsymbol{\omega}}(\mathbf{x}) = 0.5$$



EXEMPLO

Suponha que a probabilidade de um cliente adquirir um produto por mala direta é,

$$p(evento) = \frac{1}{1 + e^{-(-1.143 + 0.452x_1 + 0.029x_2 - 0.242x_3)}}$$

x_1 é sexo (1 para feminino e 0 para masculino), x_2 é idade e x_3 é estado civil (1 para solteiro e 0 para casado).

Uma pessoa do sexo feminino, com 40 anos de idade e casada, irá adquirir o produto?

$$p(evento) = \frac{1}{1 + e^{-(-1.143 + 0.452 \times 1 + 0.029 \times 40 - 0.242 \times 0)}} = 0.61$$

Sim.

NOSSO PROBLEMA, ENTÃO É...

Conjunto de m dados treinamento: $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}$
onde

$$\mathbf{x}^{(i)} \in \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}, x_0^{(i)} = 1, \quad y \in \{0,1\}$$

Como escolho ω ????

$$h_{\omega}(\mathbf{x}) = \sigma(\omega^T \mathbf{x}) = \frac{1}{1 + e^{-\omega^T \mathbf{x}}}$$



OS PARÂMETROS ω

Os dados são usados **para treinar o modelo**, isto é, achar os parâmetros ω .

ENTENDENDO O CUSTO MATEMATICAMENTE...

Para obter a função custo, a interpretaremos estatisticamente com o método de máxima verossimilhança. Maximizar a função de verossimilhança equivale a encontrar o valor de ω que torna a observação de \mathbf{y} a **mais provável** possível,

$$\hat{\omega} = \arg \max_{\omega} p(\mathbf{y}|\mathbf{x}; \omega)$$

onde $p(\mathbf{y}|\mathbf{x}; \omega)$ é a probabilidade de todas as saídas observadas \mathbf{y} nos dados de treinamento, dadas todas as entradas \mathbf{x} e parâmetros ω . Isso determina matematicamente o que significa **mais provável**.

UM POUCO DA NOMENCLATURA A SER USADA

A probabilidade, de acordo com a definição de distribuição de Bernoulli,

$$p(y^{(i)} | \mathbf{x}^{(i)}; \boldsymbol{\omega}) = [h_{\omega}(\mathbf{x}^{(i)})]^{y^{(i)}} [1 - h_{\omega}(\mathbf{x}^{(i)})]^{1-y^{(i)}}$$

onde $h_{\omega}(\mathbf{x}^{(i)}) = \frac{1}{1+e^{-\boldsymbol{\omega}^T \mathbf{x}^{(i)}}}$

Se a resposta da classificação binária é $y^{(i)} = 1$

$$p(y^{(i)} | \mathbf{x}^{(i)}; \boldsymbol{\omega}) = h_{\omega}(\mathbf{x}^{(i)})$$

Se a resposta da classificação binária é $y^{(i)} = 0$

$$p(y^{(i)} | \mathbf{x}^{(i)}; \boldsymbol{\omega}) = 1 - h_{\omega}(\mathbf{x}^{(i)})$$

CONT...

As m observações são independentes e, portanto,

$$p(\mathbf{y}|\mathbf{x}; \boldsymbol{\omega}) = \prod_{i=1}^m p(y^{(i)}|\mathbf{x}^{(i)}; \boldsymbol{\omega}) = \prod_{i=1}^m h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})^{y^{(i)}} [1 - h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})]^{1-y^{(i)}}$$

Por razões numéricas, geralmente é melhor considerar o logaritmo de $p(\mathbf{y}|\mathbf{x}; \boldsymbol{\omega})$

$$\ln p(\mathbf{y}|\mathbf{x}; \boldsymbol{\omega}) = \sum_{i=1}^m y^{(i)} \ln h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln[1 - h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})]$$

Portanto, achar o valor mais provável equivale a

$$\hat{\boldsymbol{\omega}} = \arg \max_{\boldsymbol{\omega}} \sum_{i=1}^m y^{(i)} \ln h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln[1 - h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})]$$

FUNÇÃO PERDA (LOSS FUNCTION)

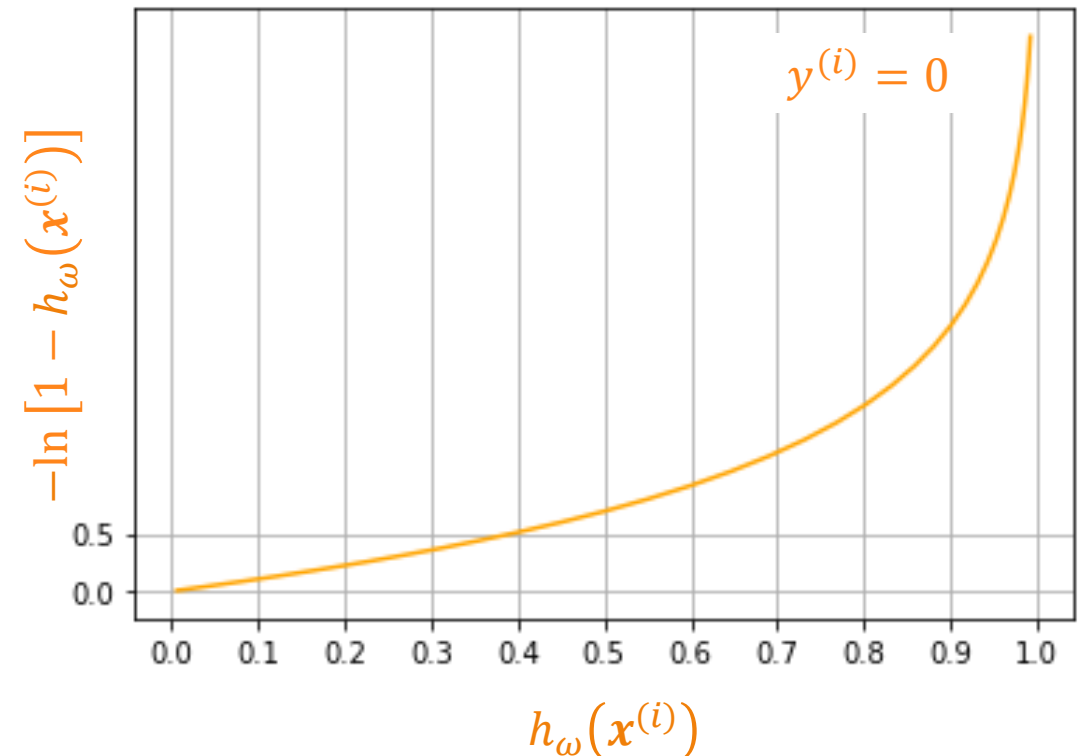
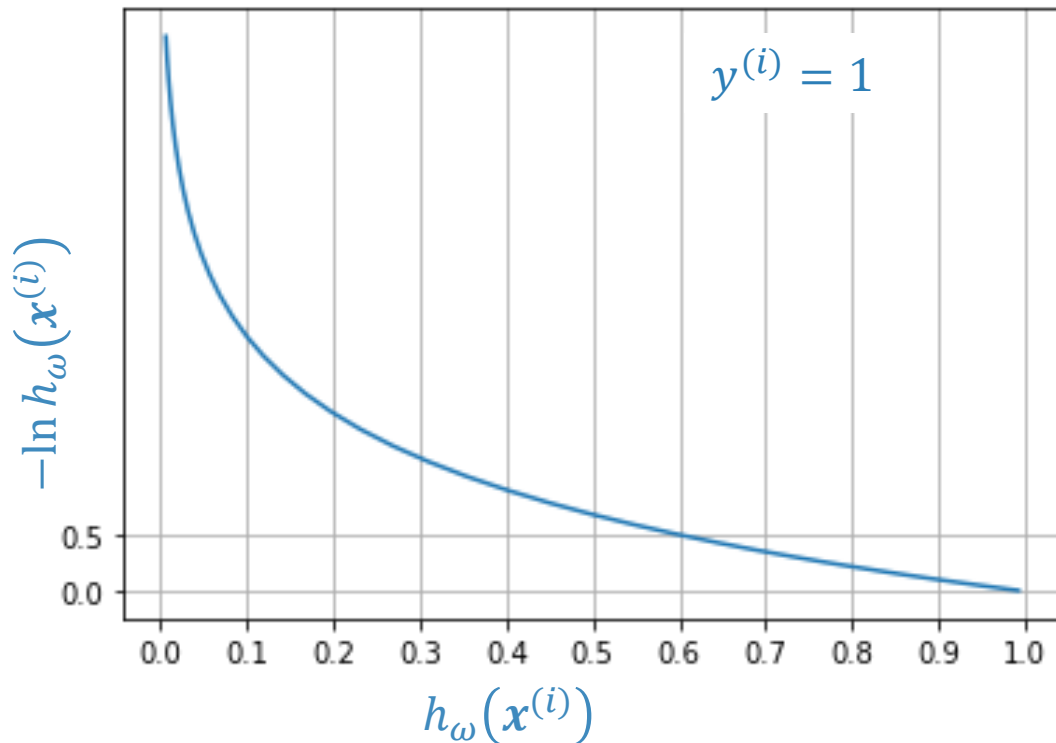
$$\ln p(y^{(i)} | \mathbf{x}^{(i)}; \boldsymbol{\omega}) = y^{(i)} \ln h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln[1 - h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})] = -L[y^{(i)}, h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})]$$

Essa função é negativa porque quando treinamos precisamos **maximizar a probabilidade minimizando o somatório da função de perda para todas as entradas, isto é, diminuindo a função custo.**

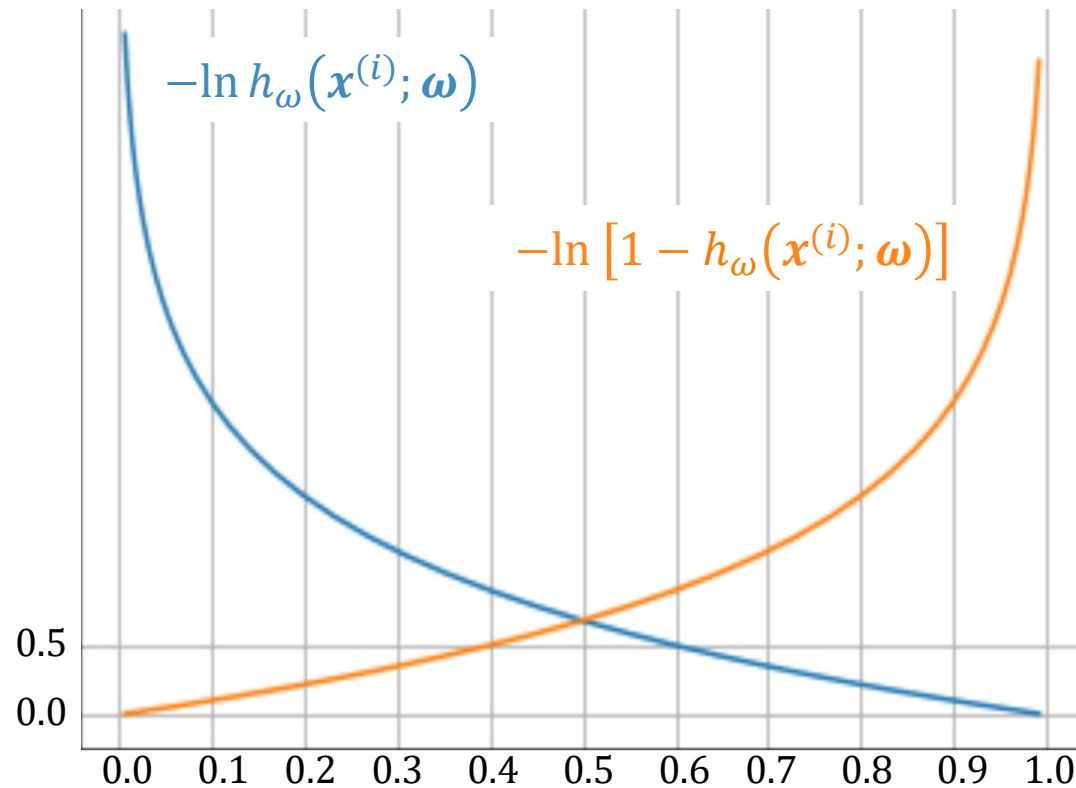
OLHANDO INICIALMENTE PARA FUNÇÃO PERDA...

$$-\{y^{(i)} \ln h_{\omega}(x^{(i)}) + (1 - y^{(i)}) \ln [1 - h_{\omega}(x^{(i)})]\} = \begin{cases} -\ln h_{\omega}(x^{(i)}) & \text{se } y^{(i)} = 1 \\ -\ln [1 - h_{\omega}(x^{(i)})] & \text{se } y^{(i)} = 0 \end{cases}$$

$y^{(i)}$ vale 0 ou 1



OU SEJA... A FUNÇÃO PERDA



$$L(\omega) = \begin{cases} -\ln h_\omega(\mathbf{x}^{(i)}; \omega) & \text{se } y^{(i)} = 1 \\ -\ln [1 - h_\omega(\mathbf{x}^{(i)}; \omega)] & \text{se } y^{(i)} = 0 \end{cases}$$

FUNÇÃO CUSTO

Fazendo o somatório em todo meu conjunto de dados, tenho:

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m \begin{cases} \ln[h_{\omega}(x^{(i)})] & \text{se } y^{(i)} = 1 \\ \ln[1 - h_{\omega}(x^{(i)})] & \text{se } y^{(i)} = 0 \end{cases}$$

entropia cruzada

Diminuir o custo aumentará a máxima verossimilhança, assumindo que as entradas são extraídas de uma distribuição identicamente independente.

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \ln[h_{\omega}(x^{(i)})] + (1 - y^{(i)}) \ln[1 - h_{\omega}(x^{(i)})]$$

Indica a distância entre o como o algoritmo acredita que essa distribuição deve ser e como ela realmente é.

Entropia cruzada como função perda (Cross Entropy Loss function). Também é conhecido como **perda de log** (log loss).

GRADIENTE DESCENDENTE

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \ln[h_{\omega}(\mathbf{x}^{(i)})] + (1 - y^{(i)}) \ln[1 - h_{\omega}(\mathbf{x}^{(i)})]$$

Queremos $\min_{\omega} J(\omega)$

Repetir até convergência {

$$\omega_{j+1} := \omega_j - \boxed{\alpha \frac{\partial}{\partial \omega_j} J(\omega)}$$

Regra da cadeia...

}

REGRA DA CADEIA NA DERIVADA

$$\begin{aligned}\frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) &= \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \omega_j} L\left(y^{(i)}, h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})\right) \\ \frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) &= \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \omega_j} \left[-y^{(i)} \ln[h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})] - (1 - y^{(i)}) \ln[1 - h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})] \right] \\ z^{(i)} &= \boldsymbol{\omega}^T \mathbf{x}^{(i)}, \quad h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)}) = \sigma(\mathbf{x}^{(i)}) = \frac{1}{1 + e^{-z^{(i)}}}\end{aligned}$$

$$\frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) = \frac{1}{m} \sum_{i=1}^m \frac{\partial L\left(y^{(i)}, h_{\boldsymbol{\omega}}(\mathbf{x}^{(i)})\right)}{\partial h_{\boldsymbol{\omega}}} \frac{\partial h_{\boldsymbol{\omega}}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial \omega_j}$$

Para $L(y^{(i)}, h_\omega) = -y^{(i)} \ln h_\omega(\mathbf{x}^{(i)}) - (1 - y^{(i)}) \ln[1 - h_\omega(\mathbf{x}^{(i)})]$

$$\frac{\partial L(y^{(i)}, h_\omega)}{\partial h_\omega} = \frac{\partial}{\partial h_\omega} [-y^{(i)} \ln[h_\omega] - (1 - y^{(i)}) \ln[1 - h_\omega]] = -y^{(i)} \frac{1}{h_\omega} - (-1)(1 - y^{(i)}) \frac{1}{1 - h_\omega}$$

Para $h_\omega(\mathbf{x}^{(i)}) = \sigma(\mathbf{x}^{(i)}) = \frac{1}{1 + e^{-z^{(i)}}}$

$$\frac{\partial h_\omega}{\partial z^{(i)}} = \frac{-(-1)e^{-z^{(i)}}}{(1 + e^{-z^{(i)}})^2} = h_\omega(1 - h_\omega)$$

e para $z^{(i)} = \boldsymbol{\omega}^T \mathbf{x}^{(i)}$

$$\frac{\partial z^{(i)}}{\partial \omega_j} = x_j^{(i)}$$

$$\frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) = \frac{1}{m} \sum_{i=1}^m \left[-y^{(i)} \frac{1}{h_\omega} + (1 - y^{(i)}) \frac{1}{1 - h_\omega} \right] [h_\omega(1 - h_\omega)] x_j^{(i)}$$

**RESOLVIDAS
TODAS AS
DERIVADAS!**



$$\frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) = \frac{1}{m} \sum_{i=1}^m \left[-y^{(i)} \frac{1}{h_{\omega}} + (1 - y^{(i)}) \frac{1}{1 - h_{\omega}} \right] [h_{\omega} (1 - h_{\omega})] x_j^{(i)}$$

$$\frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} (1 - h_{\omega}) - (1 - y^{(i)}) h_{\omega}] x_j^{(i)}$$

$$\frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} - h_{\omega}] x_j^{(i)}$$

$$\frac{\partial}{\partial \omega_j} J(\boldsymbol{\omega}) = \frac{1}{m} \sum_{i=1}^m [h_{\omega}(\mathbf{x}^{(i)}) - y^{(i)}] x_j^{(i)}$$

Controle
convergência
plotando gráfico
de J pelo número
de iterações

Algoritmo

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \ln[h_{\omega}(\mathbf{x}^{(i)})] + (1 - y^{(i)}) \ln[1 - h_{\omega}(\mathbf{x}^{(i)})]$$

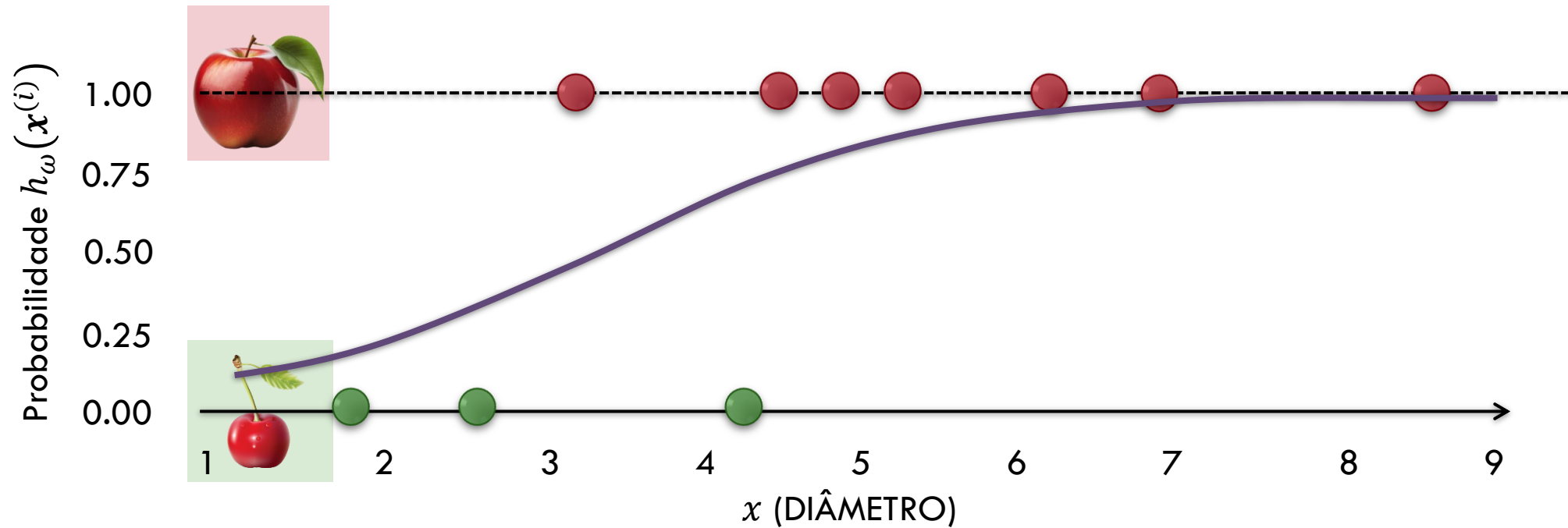
Queremos $\min_{\omega} J(\omega)$

Repetir até convergência {

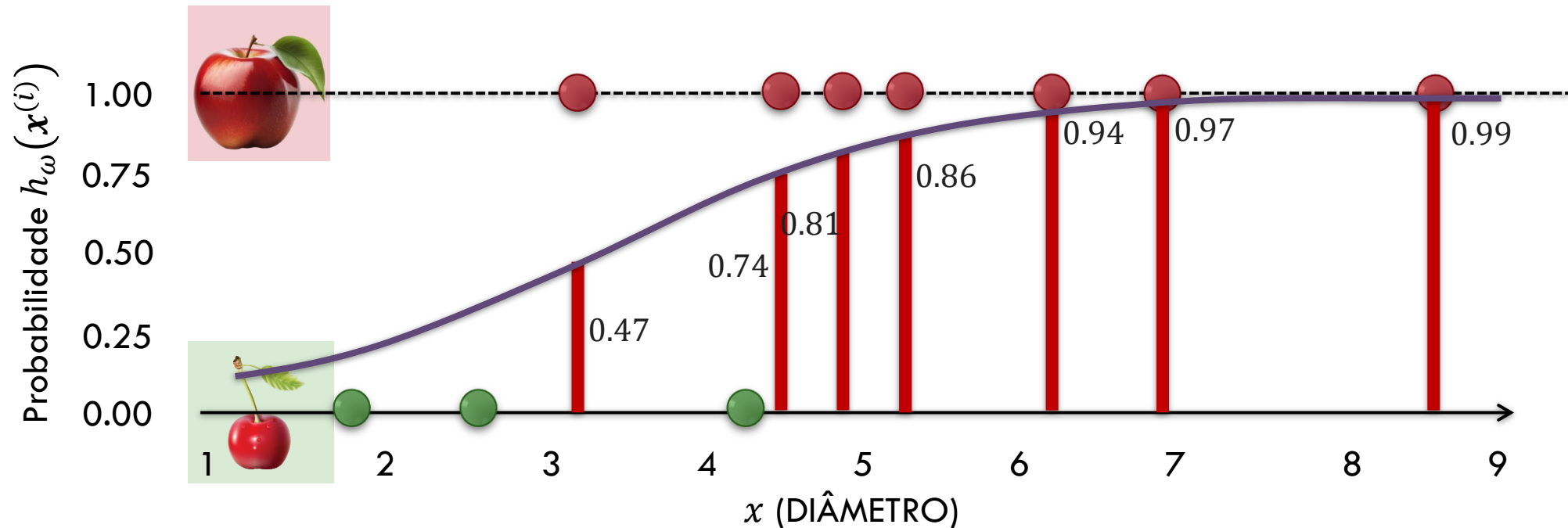
$$\omega_{j+1} := \omega_j - \alpha \frac{1}{m} \sum_{i=1}^m [h_{\omega}(\mathbf{x}^{(i)}) - y^{(i)}] \mathbf{x}_j^{(i)}$$

}

EM NOSSO EXEMPLO — MAÇÃ OU CEREJA?

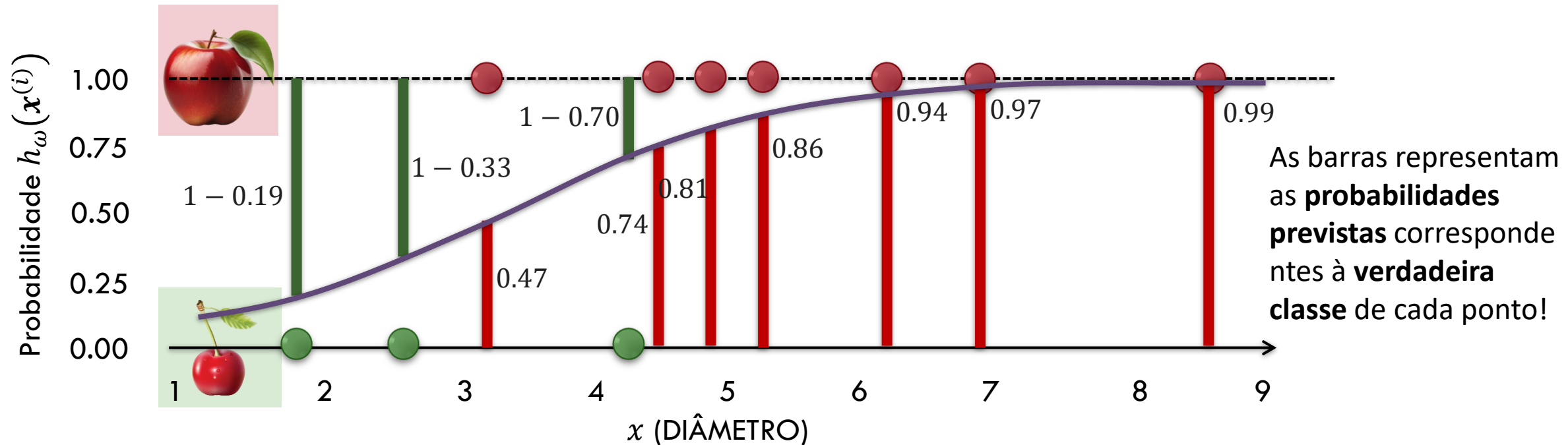


EM NOSSO EXEMPLO — MAÇÃ OU CEREJA?



Para os pontos que pertencem à **classe I (maçã)**, quais as **probabilidades** previstas pelo nosso classificador? Essas são as **barras vermelhas** abaixo da **curva sigmoide**, posicionadas sobre o atributo x de cada um dos pontos.

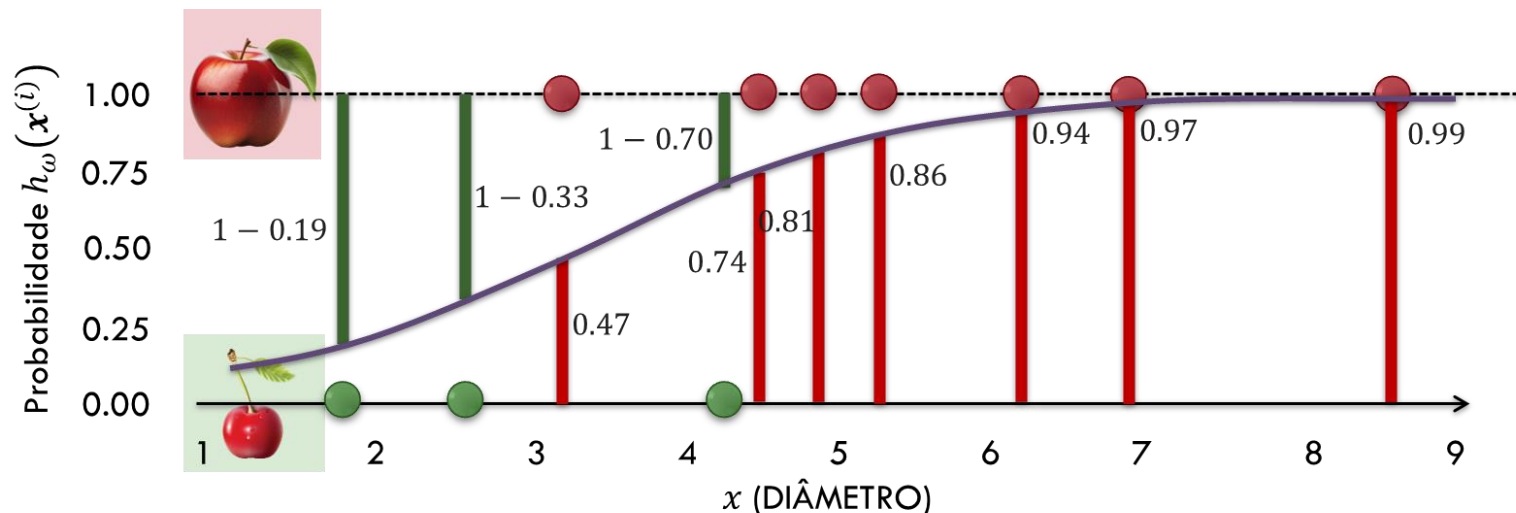
EM NOSSO EXEMPLO – MAÇÃ OU CEREJA?



Mas e os pontos da **classe 0 (cereja)**? Ora, se as **barras vermelhas** abaixo da curva sigmoide representam as probabilidades dos pontos serem **vermelhos**, como seriam as probabilidades dos pontos representarem a **classe 0**?

Barras verdes ACIMA da curva sigmoide!

E QUAL A PERDA DE UM MODELO QUE MOSTRE O DESEMPENHO APRESENTADO?



$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \ln[h_{\omega}(x^{(i)})] + (1 - y^{(i)}) \ln[1 - h_{\omega}(x^{(i)})]$$

$$J(\omega) = -\frac{1}{10} [\ln(1 - 0.19) + \ln(1 - 0.33) + \ln(0.47) + \ln(1 - 0.7) + \ln(0.74) + \ln(0.81) + \ln(0.86) + \ln(0.94) + \ln(0.97) + \ln(0.99)]$$

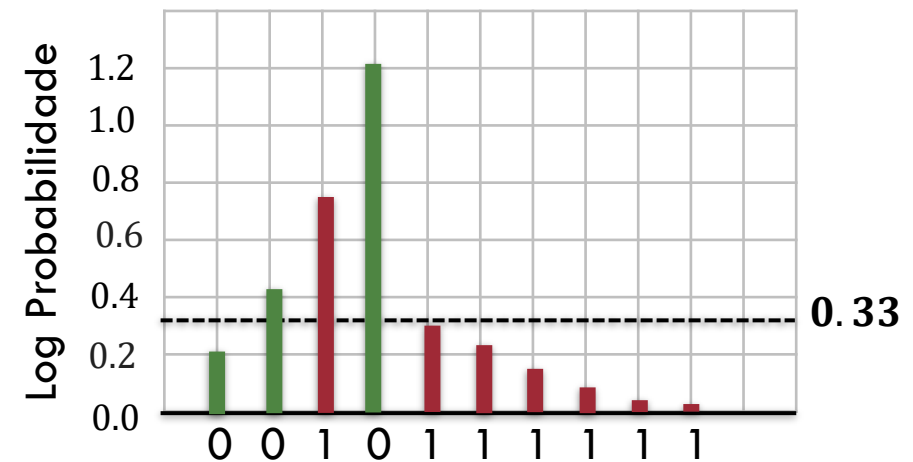
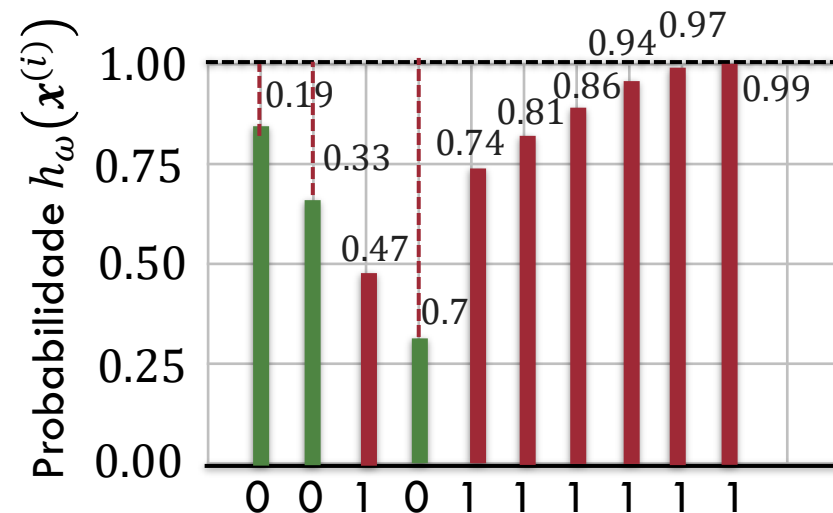
$$J(\omega) = 0.3329$$

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \ln[h_{\omega}(x^{(i)})] + (1 - y^{(i)}) \ln[1 - h_{\omega}(x^{(i)})]$$

$$J(\omega) = -\frac{1}{10} \left[\begin{array}{c} \ln(1 - 0.19) + \ln(1 - 0.33) + \ln(0.47) + \ln(1 - 0.7) + \ln(0.74) + \\ \ln(0.81) + \ln(0.86) + \ln(0.94) + \ln(0.97) + \ln(0.99) \end{array} \right] = 0.3329$$

Quão próxima é a distribuição prevista da distribuição verdadeira?

É isso que determina o erro de entropia cruzada.





MÉTRICAS

O sucesso ou fracasso dos modelos de aprendizado de máquina depende de como avaliamos seu desempenho.

MÉTRICAS

Matriz de confusão é uma medida de desempenho para o problema de classificação de aprendizado de máquina em que a saída pode ser duas ou mais classes.

Previsão↓ Entrada→	Positivo (1)	Negativo(0)
Positivo (1)	VP	FP
Negativo (0)	FN	VN

Acurácia

$$A = \frac{VP + VN}{VP + VN + FP + FN}$$

performance geral do modelo

Acurácia indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente. A acurácia é uma boa indicação geral do desempenho do modelo. Porém, pode haver situações em que ela é enganosa. Não nos fornece nenhuma informação específica da classe, como quais limites de classe foram bem aprendidos, onde o modelo foi mais confuso, etc.

Previsão↓ Entrada→	Positivo (1)	Negativo(0)
Positivo (1)	VP	FP
Negativo (0)	FN	VN

Acurácia

$$A = \frac{VP + VN}{VP + VN + FP + FN}$$

performance geral do modelo

Precisão

$$P = \frac{VP}{VP + FP}$$

Diagram illustrating the precision formula. The numerator is VP (True Positives) and the denominator is $VP + FP$ (True Positives + False Positives). Arrows point from the text '70' and '100' to the numerator and denominator respectively.

dentre as classificações positivas que o modelo fez, quantas estão realmente corretas

70

100

A precisão tenta responder à seguinte pergunta: “Qual a proporção de identificações positivas estava correta?” Ou seja, mostra com que frequência um modelo de ML está correto ao prever a classe alvo.

Se um modelo classificou um total de 100 amostras como sendo de classe positiva, e 70 delas realmente pertenciam à classe positiva do conjunto de dados (e 30 eram amostras de classe negativa previstas incorretamente como “positivas” pelo classificador), então a precisão é de 70%.

Se nosso modelo tem uma precisão de 0,7 significa que, quando prevê a classe 1, está correto em 70% do tempo.

Previsão↓ Entrada→	Positivo (1)	Negativo(0)
Positivo (1)	VP	FP
Negativo (0)	FN	VN

ABORDAGEM FOCADA EM **PRECISÃO**: IDENTIFICADOR DE IMPRESSÃO DIGITAL NA CIA

Uma previsão *positiva* é quando rotula uma pessoa como conhecida e permite acesso.

Falsos positivos serão um desastre. Alguém recebe autoridade para fazer algo que não está autorizado...

Falsos negativos são um problema menor, você não é um cliente, é um empregado...

Em uma abordagem focada na precisão, a prioridade é garantir que todas as pessoas com acesso realmente mereçam esse rótulo, **mesmo que isso signifique transtorno a alguns funcionários que não sejam detectados.**

Em essência, uma abordagem focada na precisão envolve discernimento – é preferível ignorar alguns funcionários do que arriscar classificar suspeitos incorretamente.

$$P = \frac{\text{Verdadeiros Positivos}}{\text{N.de positivos previstos}} = \frac{VP}{VP+FP}$$

A precisão é uma métrica útil nos casos em que os **Falsos Positivos** são uma preocupação maior do que os **Falsos Negativos**.



CIA verifica a impressão digital para segurança.

Acurácia

$$A = \frac{VP + VN}{VP + VN + FP + FN}$$

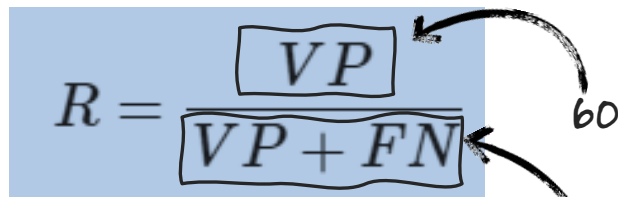
performance geral do modelo

Precisão

$$P = \frac{VP}{VP + FP}$$

dentre as classificações positivas que o modelo fez, quantas estão realmente corretas

Recall (Revocação)

$$R = \frac{VP}{VP + FN}$$


ou sensibilidade, dentre todas as situações de classe positiva como valor esperado, quantas estão corretas

Recall, por outro lado, concentra-se em capturar todos os aspectos positivos reais. Responde à pergunta: "Que proporção de positivos reais foi identificada corretamente?"

Ou seja, se o conjunto de teste de um conjunto de dados consiste em 100 amostras em sua classe positiva, quantas delas foram identificadas? Se 60 das amostras positivas foram identificadas corretamente, então o recall é de 60%.

Previsão↓ Entrada→	Positivo (1)	Negativo(0)
Positivo (1)	VP	FP
Negativo (0)	FN	VN

ABORDAGEM FOCADA EM **RECALL**: IDENTIFICADOR DE IMPRESSÃO DIGITAL NO MERCADO

Uma previsão *positiva* é quando rotula uma pessoa como conhecida e permite acesso.

Falsos negativos custarão caro ao supermercado, porque os clientes ficarão bem chateados...

Falsos positivos são um problema menor, você deu desconto a quem não deveria...

Uma abordagem focada em Recall visa sinalizar corretamente todas as pessoas que merecem o desconto, minimizando o risco de incomodar um cliente importante.

O lema aqui é: «Garantir que ninguém que merece fique sem desconto, mesmo ao custo de alguns descontos extra.»

$$R = \frac{\text{Verdadeiros Positivos}}{\text{N.de positivos previstos}} = \frac{VP}{VP+FN}$$

Recall é uma métrica útil nos casos em que o **Falsos Negativos** são considerados mais prejudiciais que os **Falsos Positivos**.



Supermercado verifica a impressão digital para descontos.

Acurácia

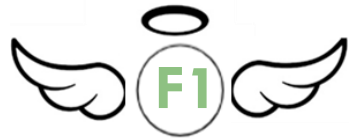
$$A = \frac{VP + VN}{VP + VN + FP + FN}$$

performance geral do modelo

Precisão

$$P = \frac{VP}{VP + FP}$$

dentre as classificações positivas que o modelo fez, quantas estão realmente corretas



$$F1 = \frac{2PR}{P + R}$$

a métrica harmonizadora:
média harmônica
entre **precisão** e **recall**.

Recall (Revocação)

$$R = \frac{VP}{VP + FN}$$

ou sensibilidade, dentre todas as situações de classe positiva como valor esperado, quantas estão corretas

Para transcender as limitações de focar isoladamente na Precisão ou no Recall, a Pontuação F1 surge como uma métrica harmoniosa. Esta pontuação busca um equilíbrio entre Precisão e Recall, fornecendo uma medida única para otimizar para um modelo mais equilibrado.

Quando priorizar a **precisão**:

- A precisão das previsões positivas é crítica.
- A falta de alguns casos positivos é aceitável.

Quando priorizar o **recall**:

- Identificar todos os casos positivos é vital.
- Alguns falsos positivos são um compromisso aceitável.

Previsão↓ Entrada→	Positivo (1)	Negativo(0)
Positivo (1)	VP	FP
Negativo (0)	FN	VN

F1

	Precisão (P)	Recall (R)
Algoritmo 01	0.5	0.4
Algoritmo 02	0.7	0.1
Algoritmo 03	0.02	1.0



Se temos diferentes algoritmos, como comparamos diferentes valores de recall e precisão?
Podem ser 3 algoritmos diferentes, ou 3 algoritmos iguais, com diferentes valores de limites.

Como decidimos qual é o melhor?

F1

	Precisão (P)	Recall (R)	Média
Algoritmo 01	0.5	0.4	0.45
Algoritmo 02	0.7	0.1	0.4
Algoritmo 03	0.02	1.0	0.51

$$M = \frac{P + R}{2}$$

Média não é uma boa opção. Veja que no exemplo o melhor algoritmo, de acordo com a média é o Modelo 3, que tem recall =1 e, portanto, prevê y=1 o tempo todo. Ele tem uma precisão muitíssimo baixa.

F1

	Precisão (P)	Recall (R)	Média	F1
Algoritmo 01	0.5	0.4	0.45	0.444
Algoritmo 02	0.7	0.1	0.4	0.175
Algoritmo 03	0.02	1.0	0.51	0.0392

$$M = \frac{P + R}{2}$$

$$F1 = \frac{2PR}{P + R}$$

F1

	Precisão (P)	Recall (R)	Média	F1
Algoritmo 01	0.5	0.4	0.45	0.444
Algoritmo 02	0.7	0.1	0.4	0.175
Algoritmo 03	0.02	1.0	0.51	0.0392

$$M = \frac{P + R}{2}$$

$$F1 = \frac{2PR}{P + R}$$



$P = 0 \text{ ou } R = 0 \rightarrow F1 = 0$



$P = 1 \text{ e } R = 1 \rightarrow F1 = 1$

Acurácia

$$A = \frac{VP + VN}{VP + VN + FP + FN}$$

performance geral do modelo

Precisão

$$P = \frac{VP}{VP + FP}$$

dentre as classificações positivas que o modelo fez, quantas estão realmente corretas

Recall (Revocação)

$$R = \frac{VP}{VP + FN}$$

ou sensibilidade, dentre todas as situações de classe positiva como valor esperado, quantas estão corretas

Taxa de Falsos Positivos

$$FPR = \frac{FP}{FP + VN}$$

também é chamada de taxa de alarme falso, pois resume a frequência com que uma classe positiva é prevista quando o resultado real é negativo.

F1

$$F1 = \frac{2PR}{P + R}$$

a métrica harmonizadora: média harmônica entre **precisão** e **recall**.

Previsão↓ Entrada→	Positivo (1)	Negativo(0)
Positivo (1)	VP	FP
Negativo (0)	FN	VN

TRADING OFF ENTRE PRECISÃO E RECALL

$$P = \frac{\text{Verdadeiros Positivos}}{\text{N. de positivos reais}} = \frac{VP}{VP + FP}$$

$$R = \frac{\text{Verdadeiros Positivos}}{\text{N. de positivos previstos}} = \frac{VP}{VP + FN}$$

$$h_{\omega}(x) = \frac{1}{1 + e^{-\omega^T x}} = \sigma(\omega^T x)$$

Prevemos 1 se $h_{\omega}(x) \geq 0.7$

Prevemos 0 se $h_{\omega}(x) < 0.7$

Supondo que queremos prever $y = 1$ somente se tivermos bastante certeza:

Alta precisão, baixo recall

TRADING OFF ENTRE PRECISÃO E RECALL

$$P = \frac{\text{Verdadeiros Positivos}}{\text{N. de positivos reais}} = \frac{VP}{VP + FP}$$

$$R = \frac{\text{Verdadeiros Positivos}}{\text{N. de positivos previstos}} = \frac{VP}{VP + FN}$$

$$h_{\omega}(x) = \frac{1}{1 + e^{-\omega^T x}} = g(\omega^T x)$$

Prevemos 1 se $h_{\omega}(x) \geq 0.3$

Prevemos 0 se $h_{\omega}(x) < 0.3$

Supondo que queremos prever $y = 1$ somente se tivermos bastante certeza:

Alta precisão, baixo recall

Supondo que queremos prever $y = 0$ somente se tivermos bastante certeza

Alto recall, baixa precisão

TRADING OFF ENTRE PRECISÃO E RECALL

$$P = \frac{\text{Verdadeiros Positivos}}{\text{N. de positivos reais}} = \frac{VP}{VP + FP}$$

$$R = \frac{\text{Verdadeiros Positivos}}{\text{N. de positivos previstos}} = \frac{VP}{VP + FN}$$

$$h_{\omega}(x) = \frac{1}{1 + e^{-\omega^T x}} = \sigma(\omega^T x)$$

Prevemos 1 se $h_{\omega}(x) \geq 0.5$

Prevemos 0 se $h_{\omega}(x) < 0.5$

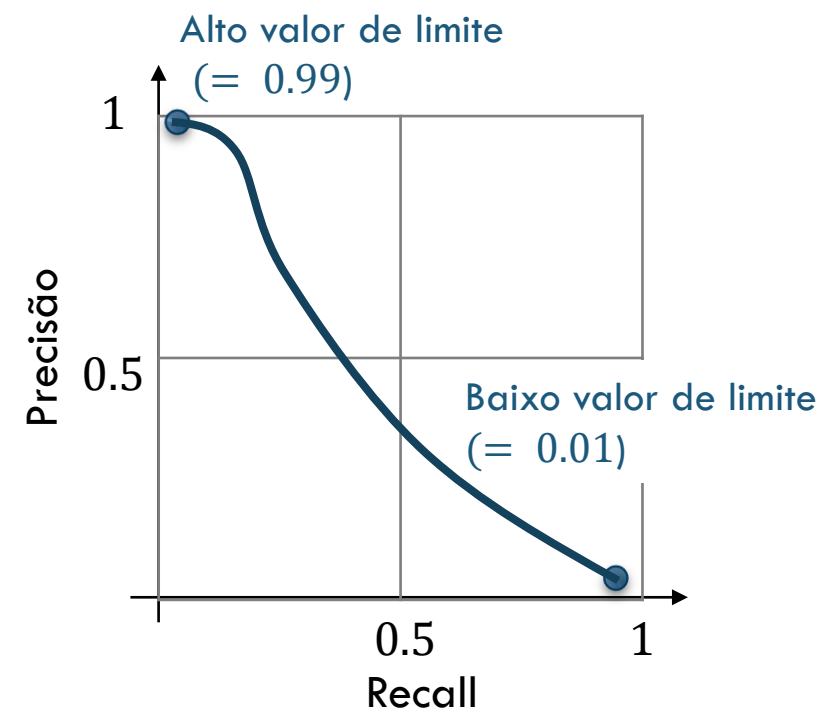
Supondo que queremos prever $y = 1$ somente se tivermos bastante certeza:

$h_{\omega}(x) \geq 0.7$: **Alta precisão, baixo recall**

Supondo que queremos prever $y = 0$ somente se tivermos bastante certeza

$h_{\omega}(x) \geq 0.3$: **Alto recall, baixa precisão**

De forma general: $h_{\omega}(x) \geq \text{limite}$





CURVA ROC RECEIVER OPERATING CHARACTERISTIC

O ROC possui dois parâmetros:
Taxa de verdadeiro positivo
Taxa de falso positivo

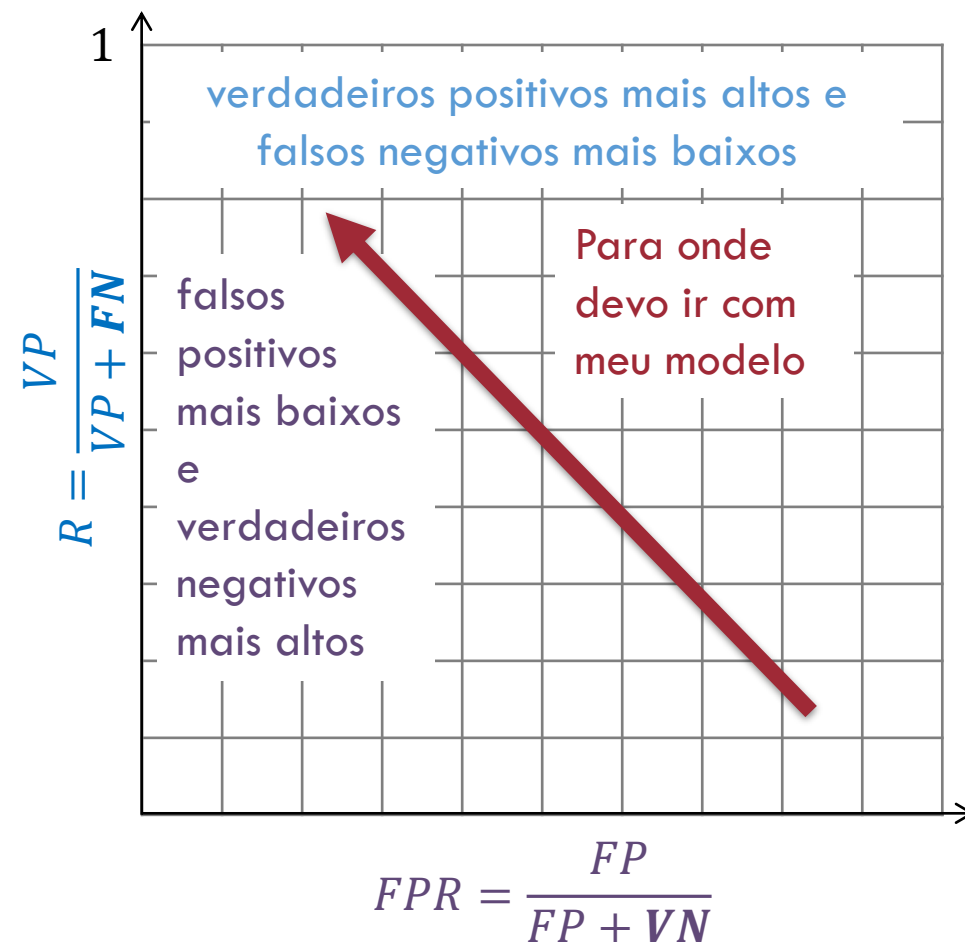
CURVA ROC E ÁREA SOB A CURVA (AUC)

A curva ROC é uma ferramenta muito útil.

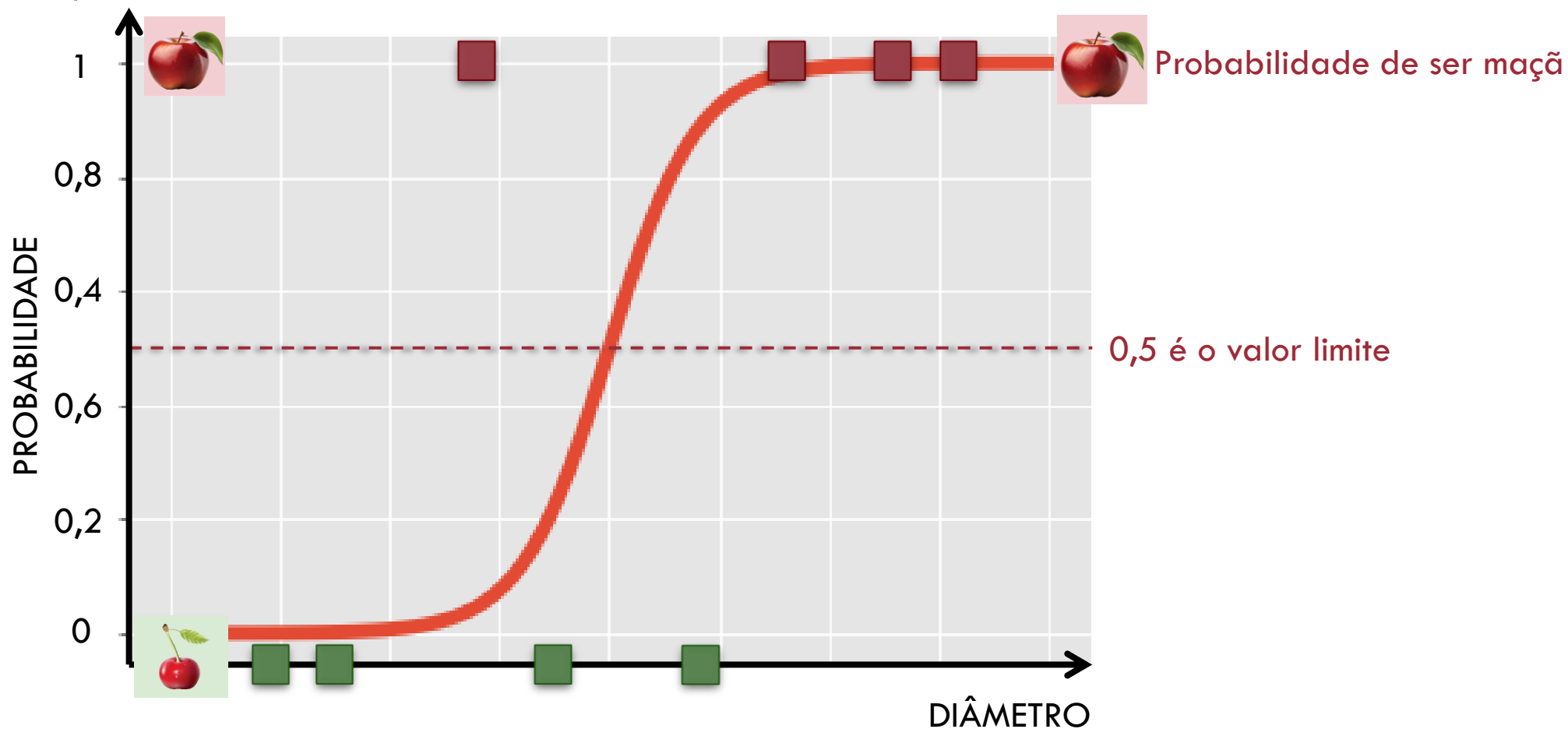
As curvas de diferentes modelos podem ser comparadas diretamente.

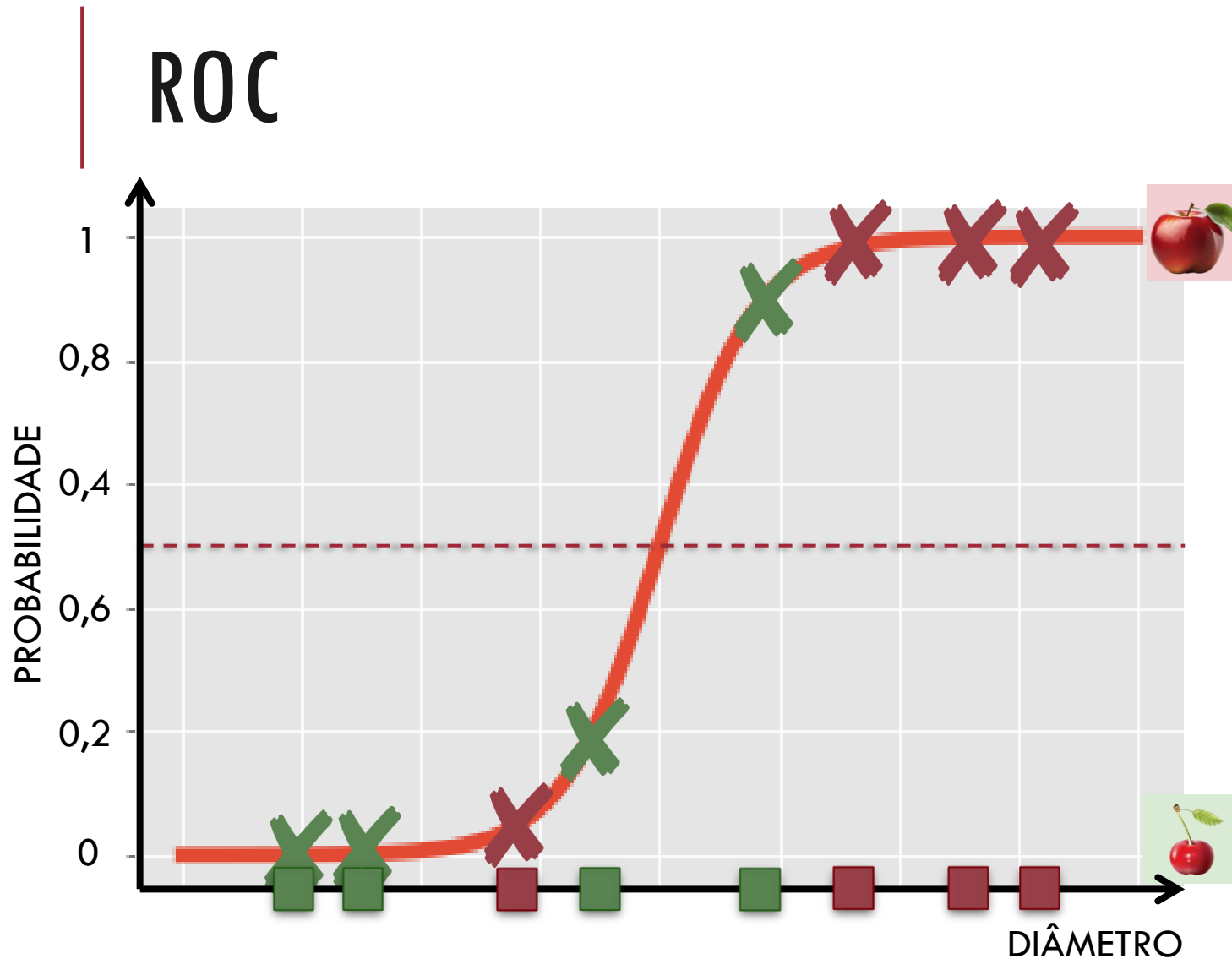
A área sob a curva (AUC) pode ser usada como um resumo da habilidade do modelo.









A forma da curva contém muitas informações, incluindo o que mais nos interessa para um problema, a taxa de falsos positivos esperada e a taxa de falsos negativos.



ROC



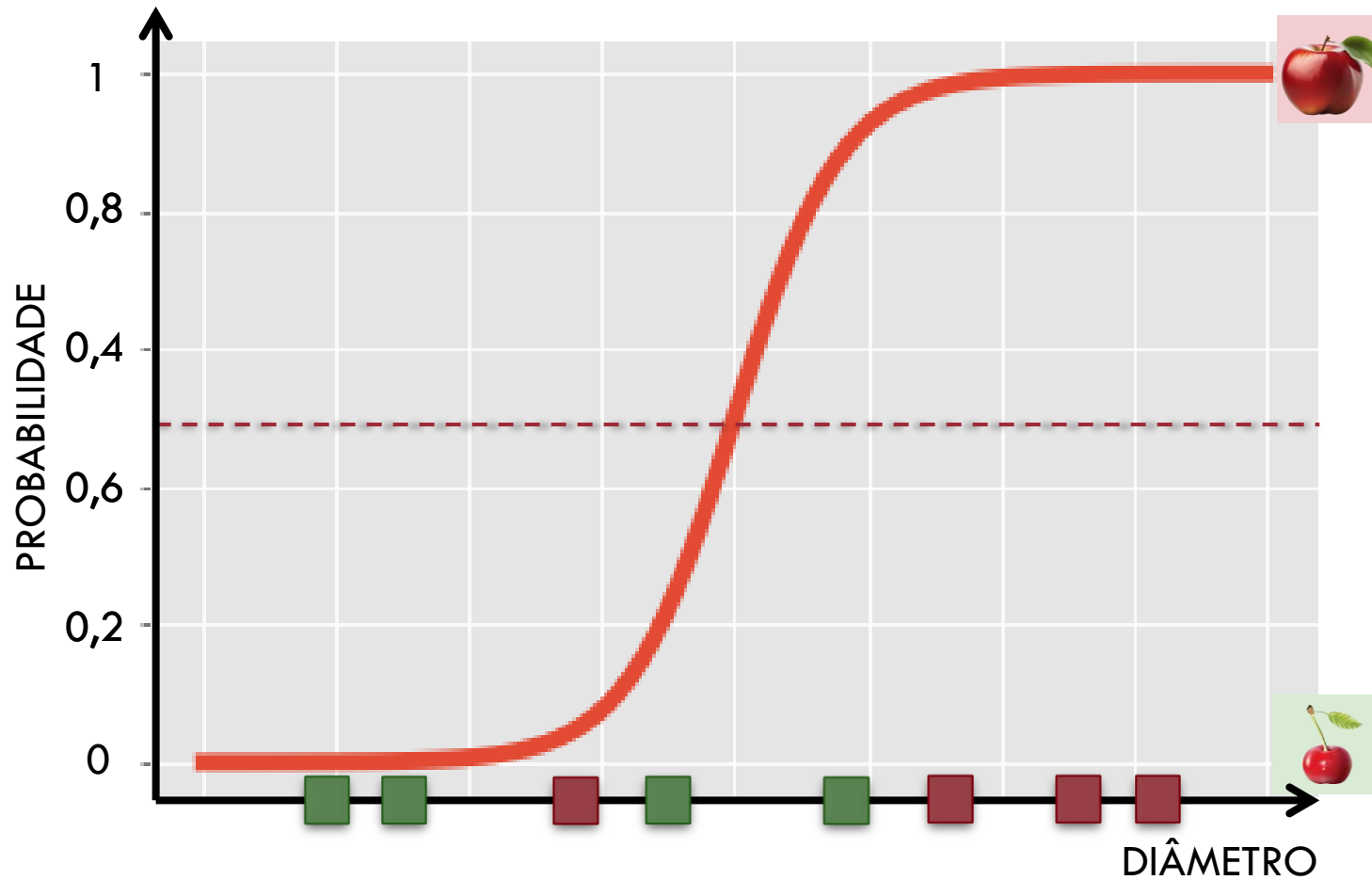


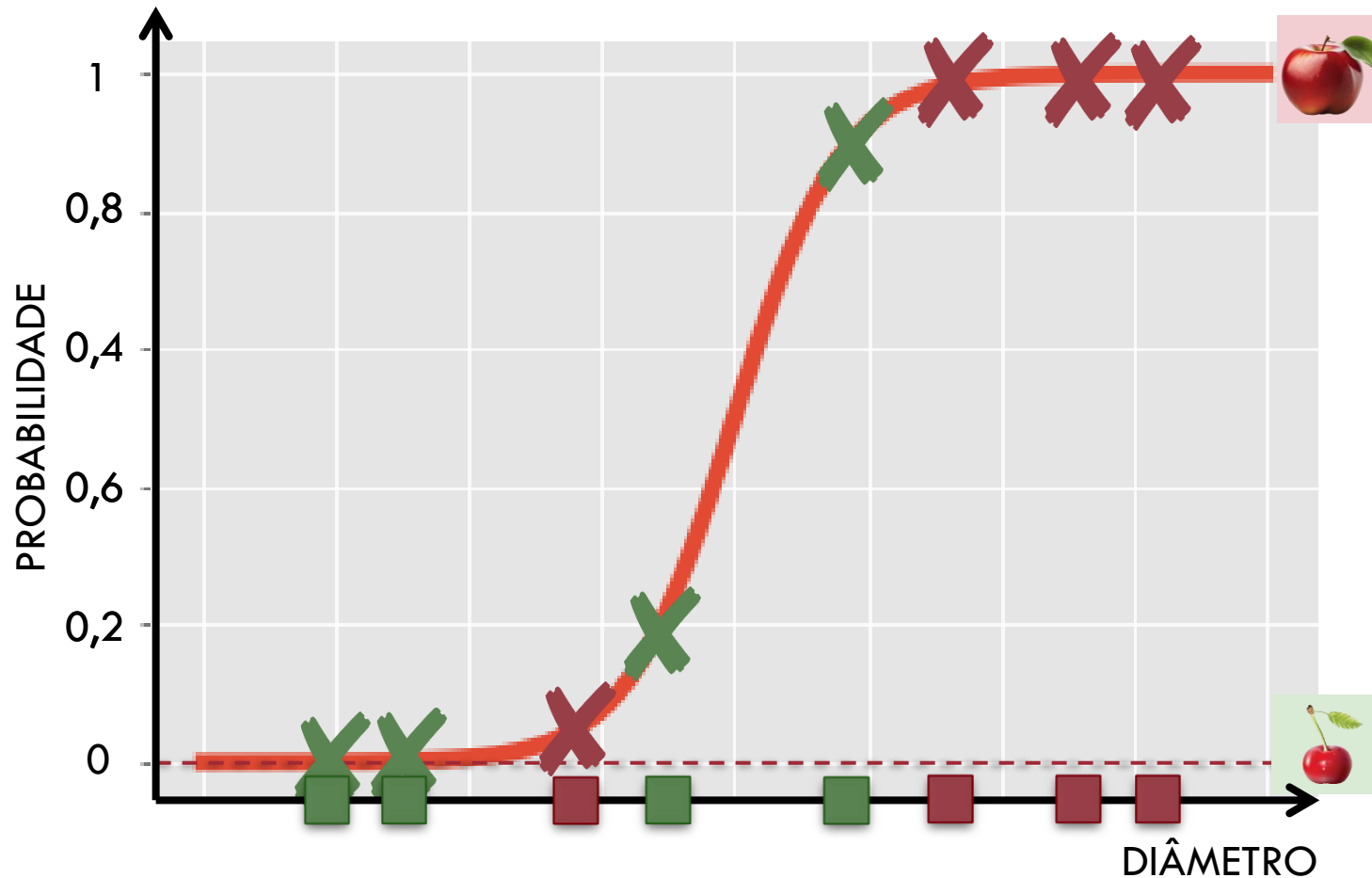
		Target		
		1 	0 	
Predicted	1 	3 	1 	FP
	0 	1 	3 	
		FN	VN	









$$R = \frac{VP}{VP + FN} = \frac{3}{4} = 0.75$$

$$FPR = \frac{FP}{FP + VN} = \frac{1}{4} = 0.25$$

E se colocarmos o limite de modo a classificarmos SEMPRE como maçã?



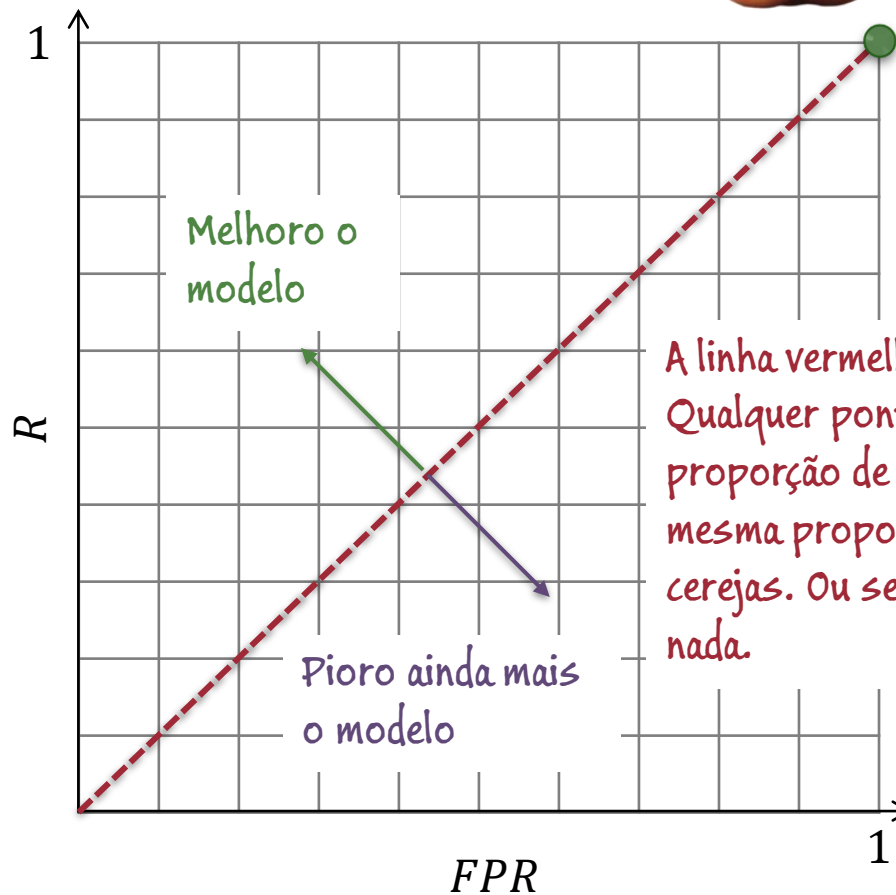


		Target		
		1 	0 	
Predicted	1 	4 	4 	FP
	0 	0 	0 	
		FN	VN	

$$R = \frac{VP}{VP + FN} = \frac{4}{4 + 0} = 1$$

$$FPR = \frac{FP}{FP + VN} = \frac{4}{4 + 0} = 1$$

ROC: R vs FPR

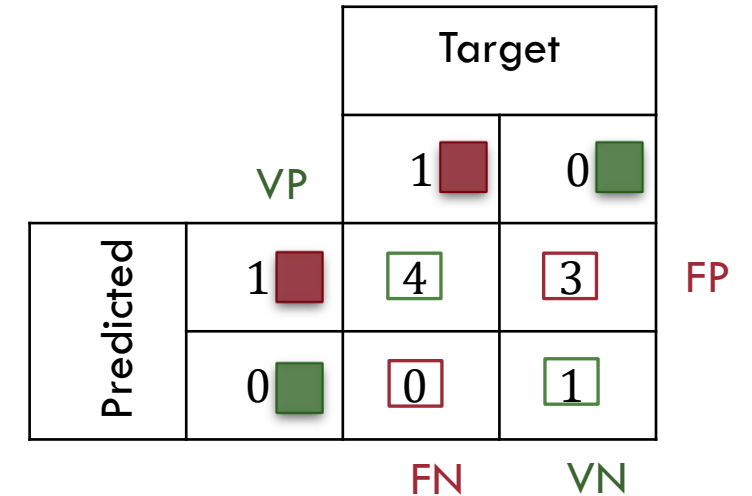


No ponto (1,1) significa que, apesar de **classificar corretamente todas as maçãs corretamente, classificou incorretamente todas as cerejas.**



Mas eu não sou maçã...

A linha vermelha diagonal significa que $FPR = R$. Qualquer ponto nessa linha significa que a proporção de classificados maçãs corretamente é a mesma proporção que classifica incorretamente as cerejas. Ou seja, o classificador não serve para nada.

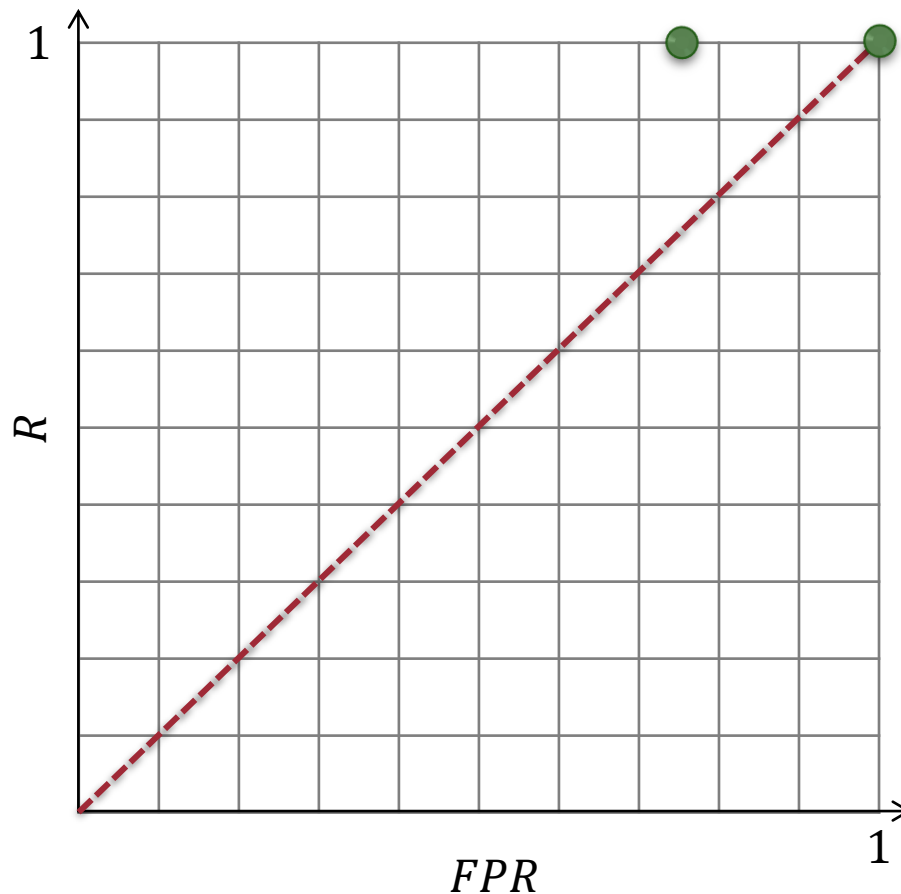


$$R = \frac{VP}{VP + FN} = \frac{4}{4 + 0} = 1$$

$$FPR = \frac{FP}{FP + VN} = \frac{3}{3 + 1} = 0.75$$

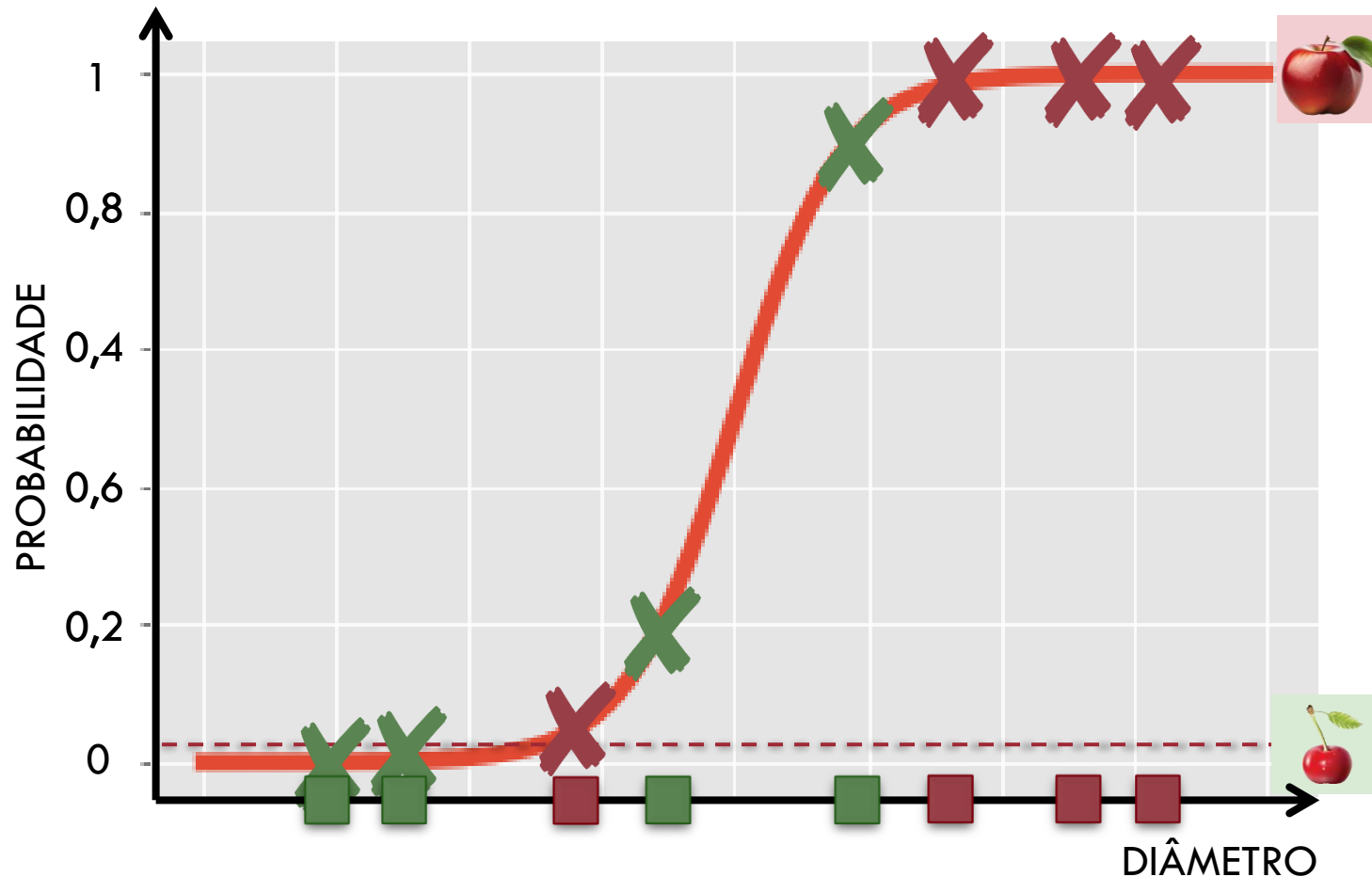
ROC: R vs FPR









O novo ponto (0.75,1) está à esquerda da linha vermelha e, portanto, sabemos que a proporção de corretamente classificados como maçãs (VP) é **maior** **que** a proporção de classificados incorretamente como maçãs (FP).



Ou seja,
O segundo ponto é melhor que o primeiro...

E se colocarmos o limite de modo que os dois menores passos são classificados como cerejas?



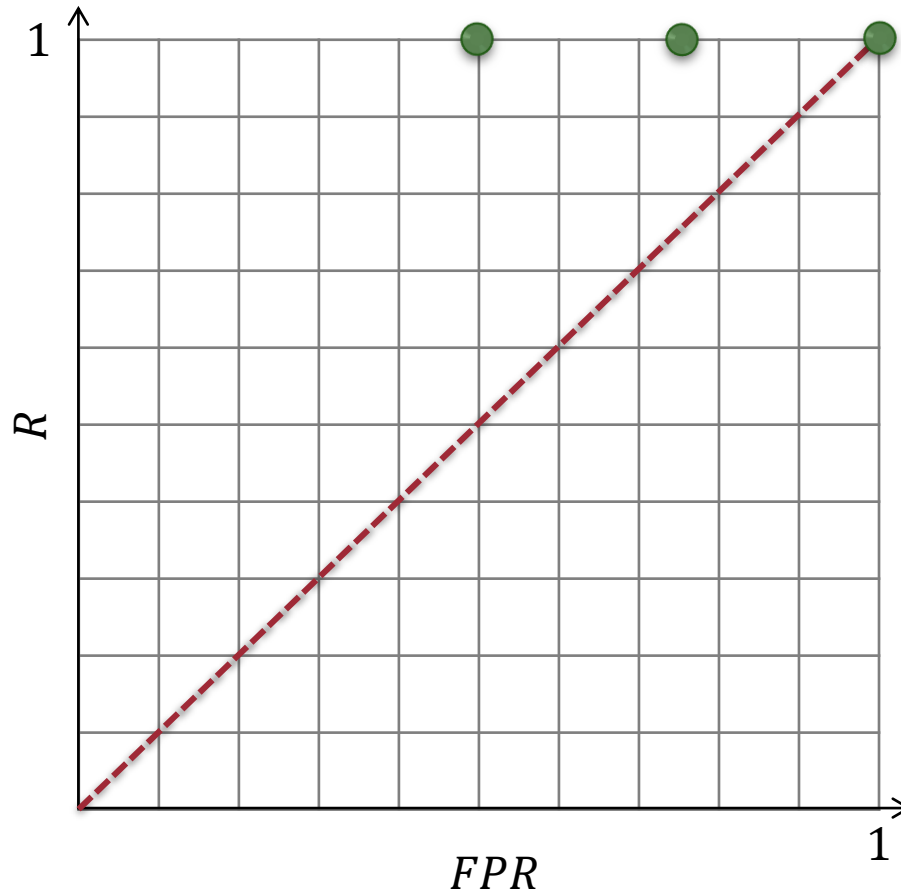
		Target		
		1 	0 	
Predicted	1 	4 	2 	FP
	0 	0 	2 	
		FN	VN	

$$R = \frac{VP}{VP + FN} = \frac{4}{4 + 0} = 1$$

$$FPR = \frac{FP}{FP + VN} = \frac{2}{2 + 2} = 0.5$$

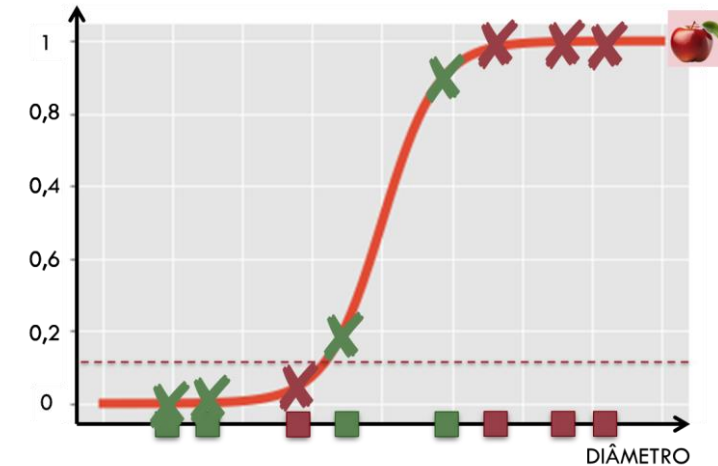
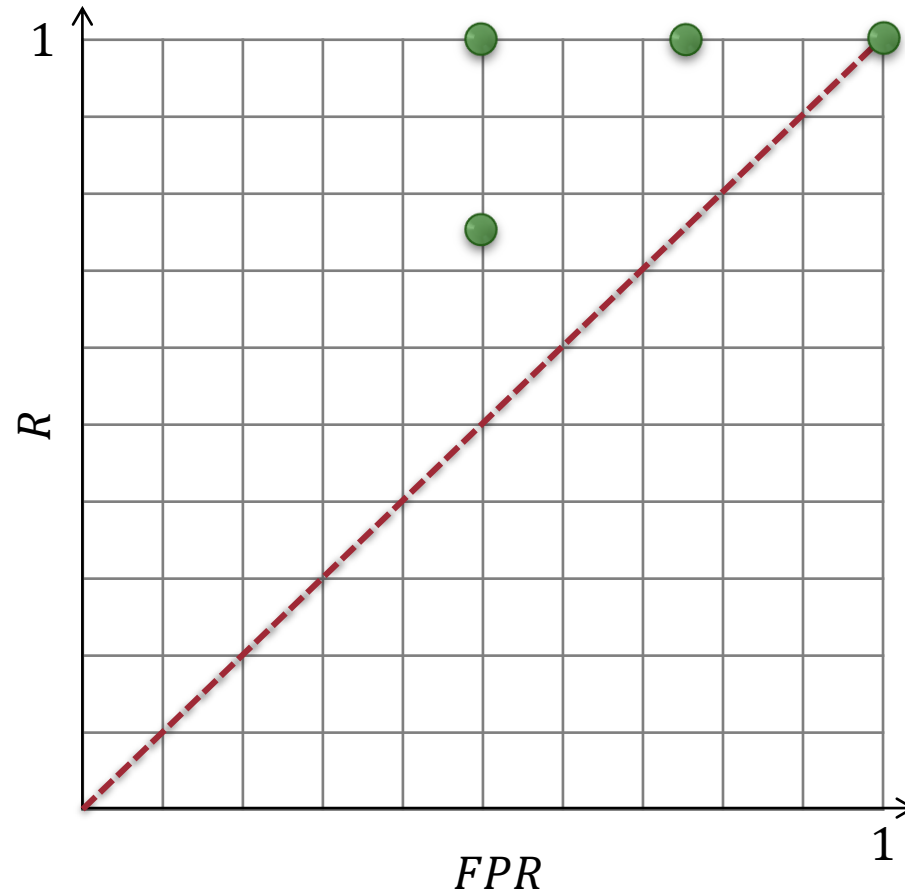
ROC: R vs FPR

O novo ponto (0.5,1) está ainda mais à esquerda da linha vermelha e, portanto, diminuiu a proporção de classificados incorretamente como instáveis (FP).



Ou seja,
O terceiro ponto é melhor que o
segundo...

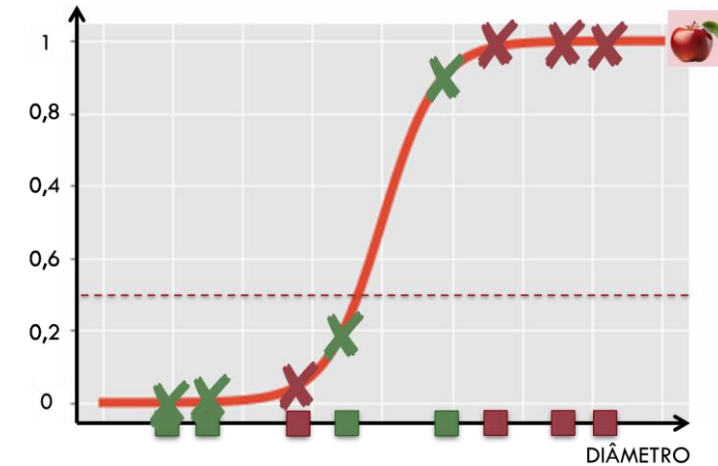
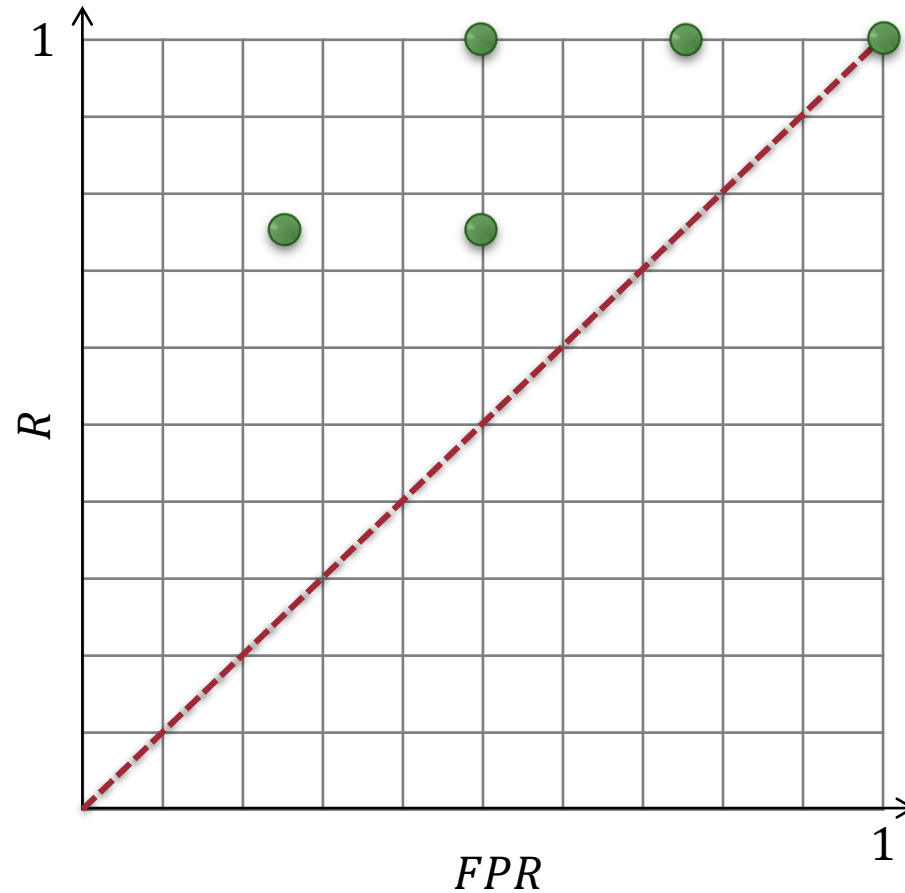
ROC: R vs FPR



$$R = \frac{VP}{VP + FN} = \frac{3}{3 + 1} = 0.75$$

$$FPR = \frac{FP}{FP + VN} = \frac{2}{2 + 2} = 0.5$$

ROC: R vs FPR

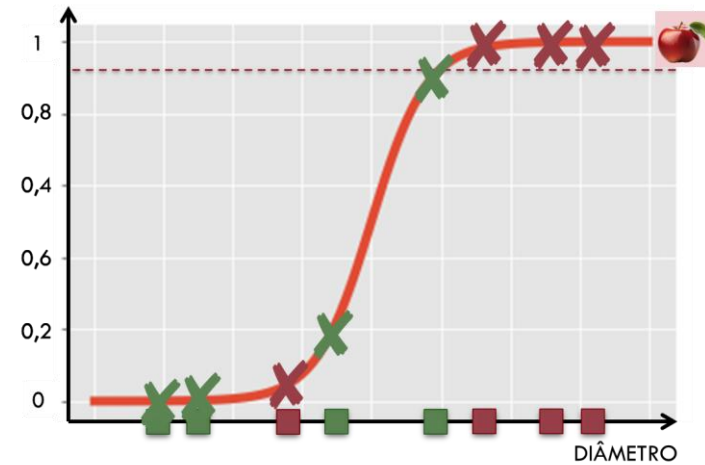
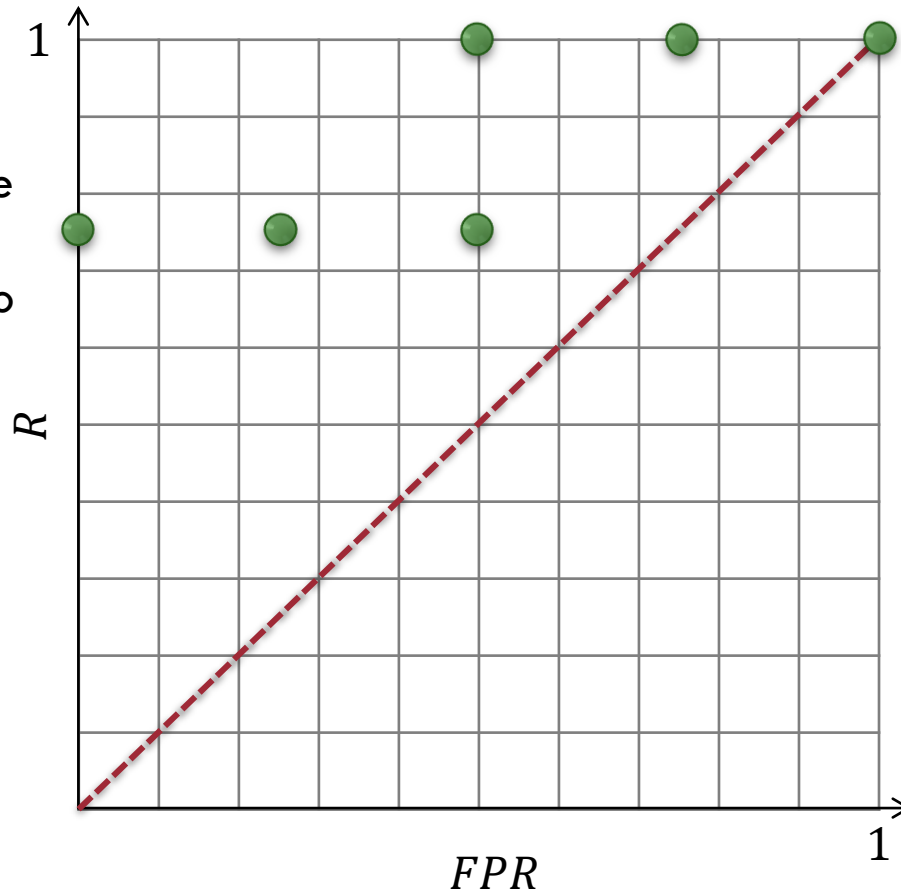


$$R = \frac{VP}{VP + FN} = \frac{3}{3 + 1} = 0.75$$

$$FPR = \frac{FP}{FP + VN} = \frac{1}{1 + 3} = 0.25$$

ROC: R vs FPR

O ponto $(0,0.75)$
classifica corretamente
75% das maçãs e
100% das cerejas. Isto
é, $FP = 0$.



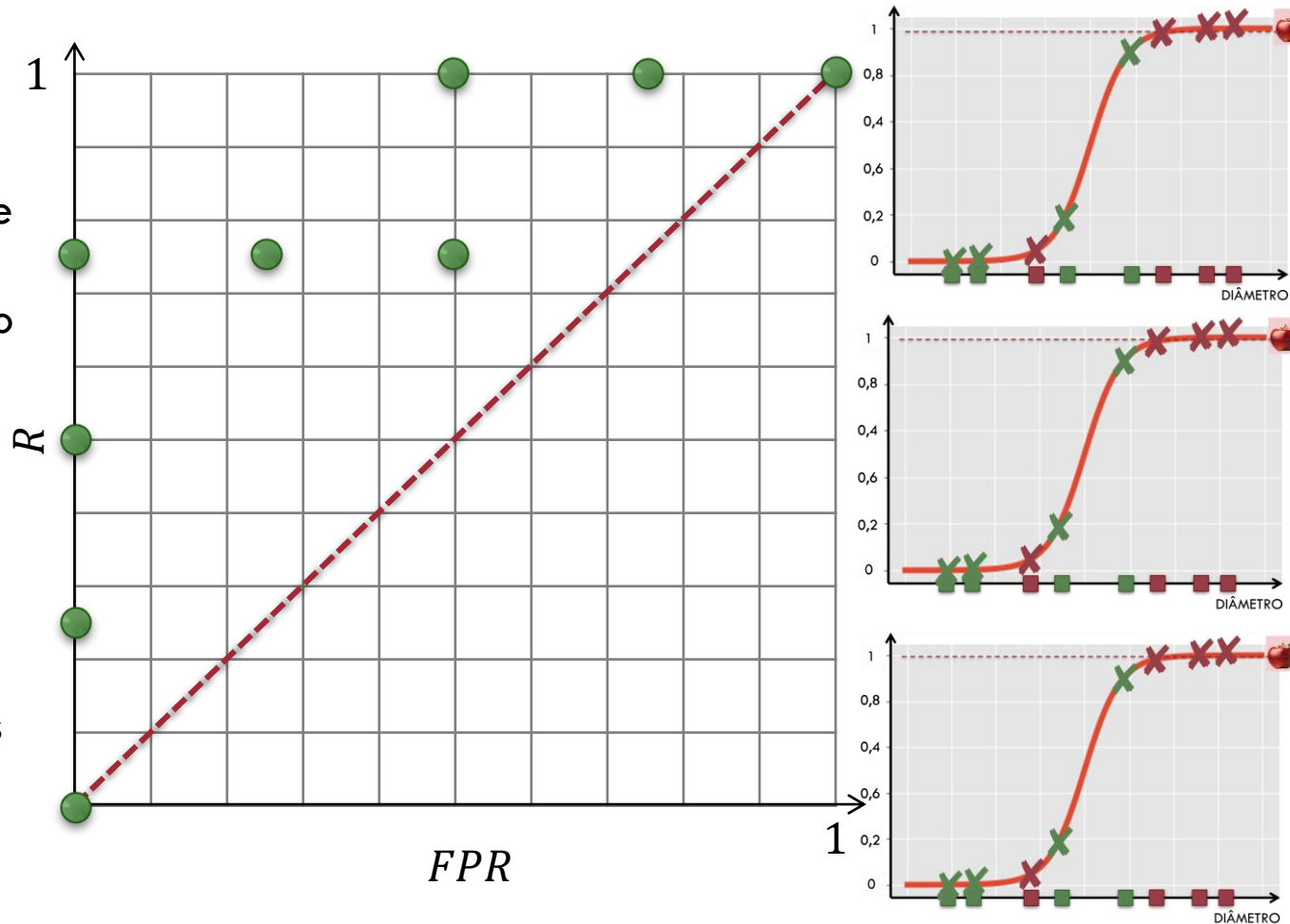
$$R = \frac{VP}{VP + FN} = \frac{3}{3 + 1} = 0.75$$

$$FPR = \frac{FP}{FP + VN} = \frac{0}{0 + 4} = 0.0$$

ROC: R vs FPR

O ponto (0,0.75)
classifica corretamente
75% das maçãs e
100% das cerejas. Isto
é, $FP = 0$.

Finalmente, escolhemos um
limite que classifique todos
os pontos como estáveis.
 $VP = 0$ e $FP = 0$.



$$R = \frac{VP}{VP + FN} = \frac{2}{2 + 2} = 0.5$$

$$FPR = 0$$

$$R = \frac{VP}{VP + FN} = \frac{1}{1 + 3} = 0.25$$

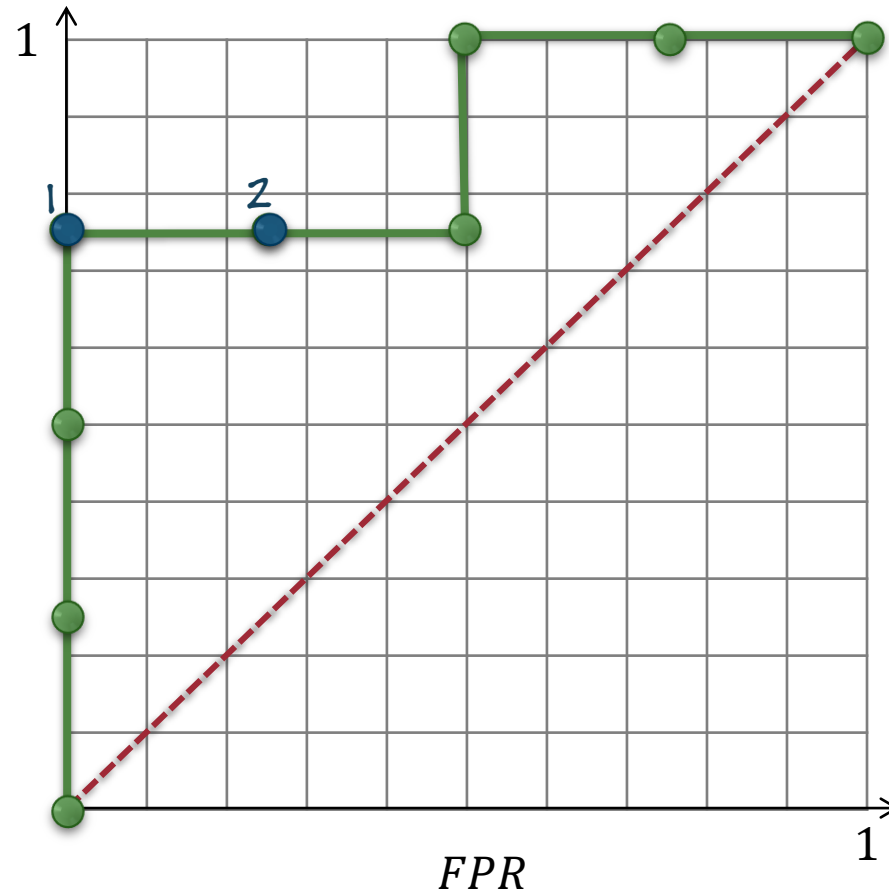
$$FPR = 0$$

$$R = \frac{VP}{VP + FN} = \frac{0}{0 + 4} = 0$$

$$FPR = 0$$

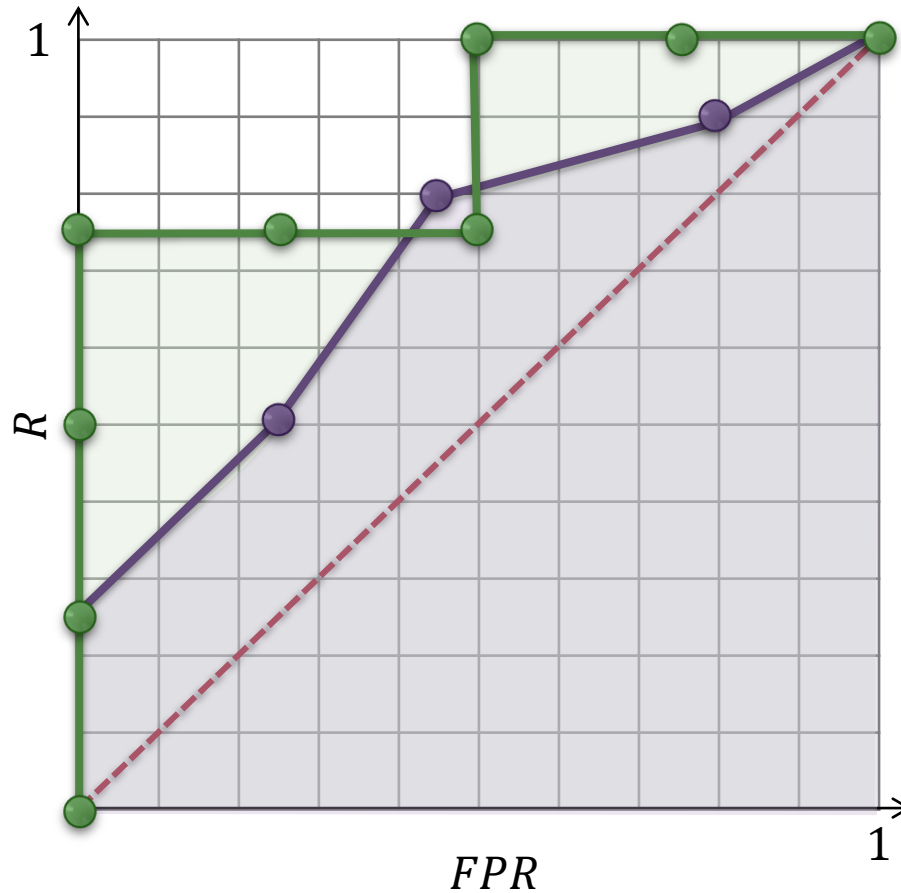
ROC: R vs FPR

Sem ter que fazer
matrizes de
confusão, limite é 1 é
melhor que 2



Curva ROC resume a matriz de
confusão de cada limite escolhido.

AUC



A AUC torna fácil a comparação entre diferentes ROC.

A curva verde tem AUC maior que a curva roxa, sugerindo que é melhor.

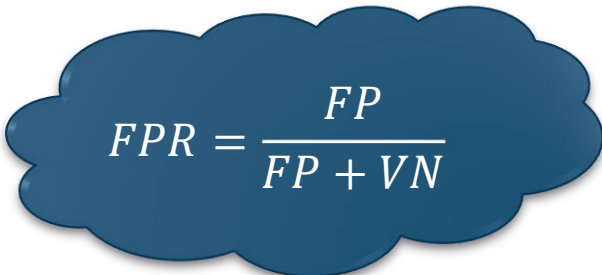
OBSERVAÇÃO

Se as amostras são desbalanceadas (por exemplo, o número de maçãs é muito maior que cerejas) então **PRECISÃO** pode ser mais útil que **FPR**.

Isso ocorre porque a precisão, definida como,

$$P = \frac{VP}{VP + FP}$$

não inclui o número de VN em seu cálculo e, portanto, não é afetada pelo desbalanceamento.


$$FPR = \frac{FP}{FP + VN}$$



REGRESSÃO LOGÍSTICA PARA MAIS DE DUAS CLASSES

OVO E OVR

Modelos de classificação binária como regressão logística e SVM não suportam a classificação multiclasse e requerem estratégias.

Uma abordagem para usar algoritmos de classificação binária para problemas de multiclassificação é dividir o conjunto de dados de classificação multiclasse em vários conjuntos de dados de classificação binária e ajustar um modelo de classificação binário em cada um.

Dois exemplos diferentes de estratégias são:

Um contra Um – *One versus One* (OvO)

Um contra o Resto – *One versus All, One versus Rest* (OvA, OvR).

ONE-VS-REST (ONE-VS-ALL)

$$y \in \{1, 2, \dots, K\}$$

Treinamos K classificadores binário separados, um para cada classe, e executamos todos esses classificadores. Cada classificador retorna um valor $[0, 1]$.

Para qualquer novo exemplo \mathbf{x} que desejamos prever escolhemos a classe com a pontuação máxima:

$$\hat{y} = \arg \max_{k \in \{1, 2, \dots, K\}} h_{\omega}^{(k)}(\mathbf{x}).$$

$h_{\omega}^{(1)}$ é um classificador binário projetado para reconhecer objetos da classe 1 entre todos os objetos das outras classes.

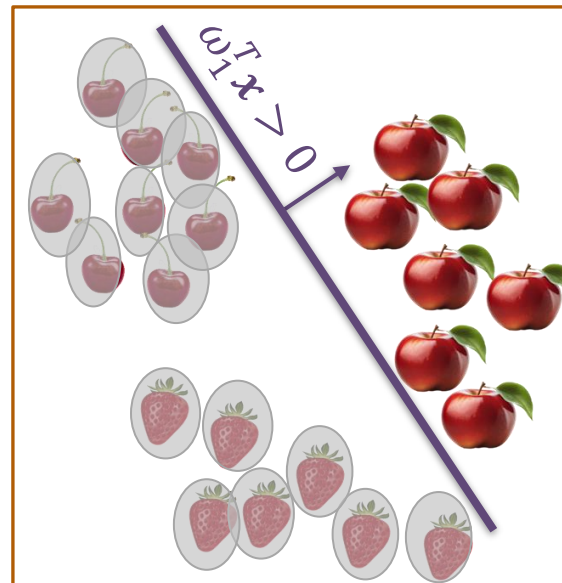
$$h_{\omega}^{(1)}(\mathbf{x}) = p(y = 1 | \mathbf{x}; \omega_1)$$

$$h_{\omega}^{(2)}(\mathbf{x}) = p(y = 2 | \mathbf{x}; \omega_2)$$

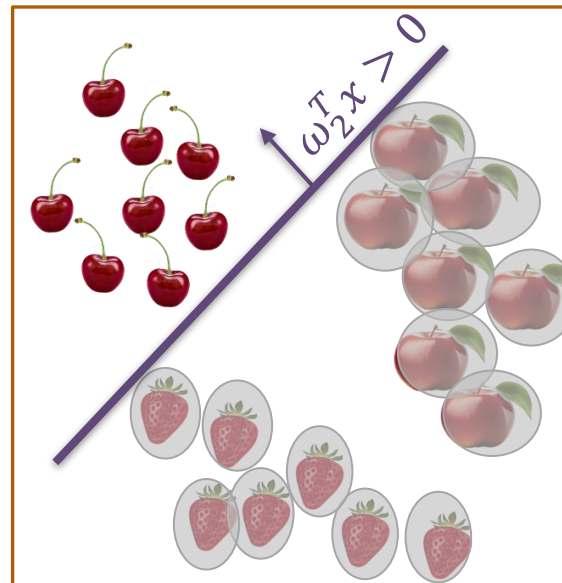
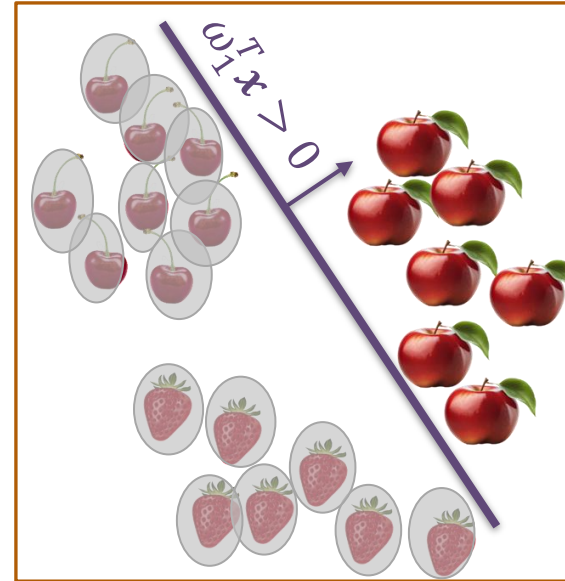
\vdots

$$h_{\omega}^{(K)}(\mathbf{x}) = p(y = K | \mathbf{x}; \omega_K)$$

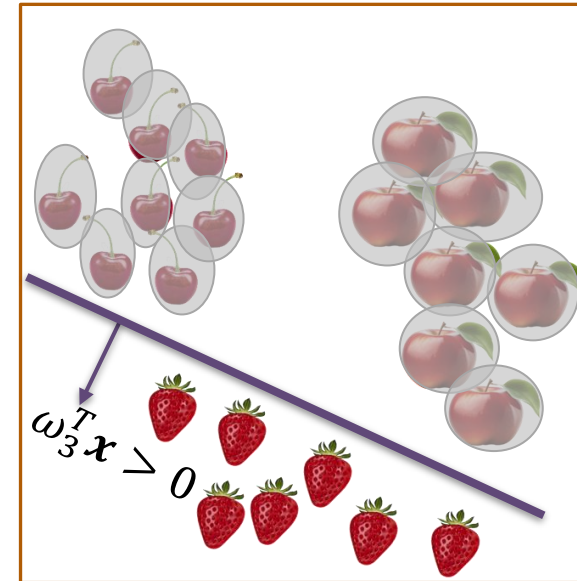
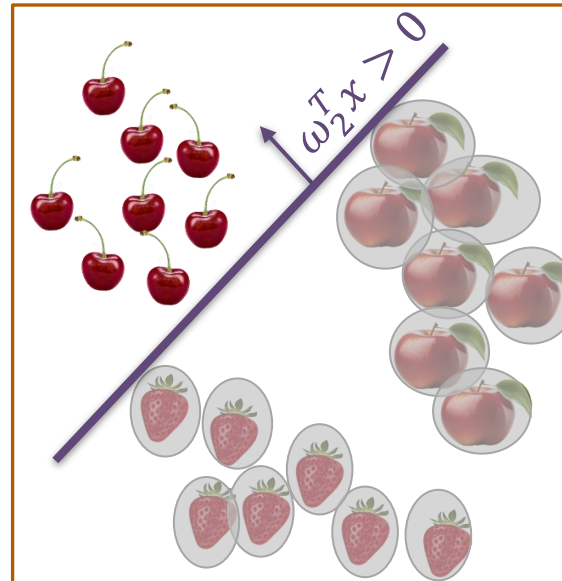
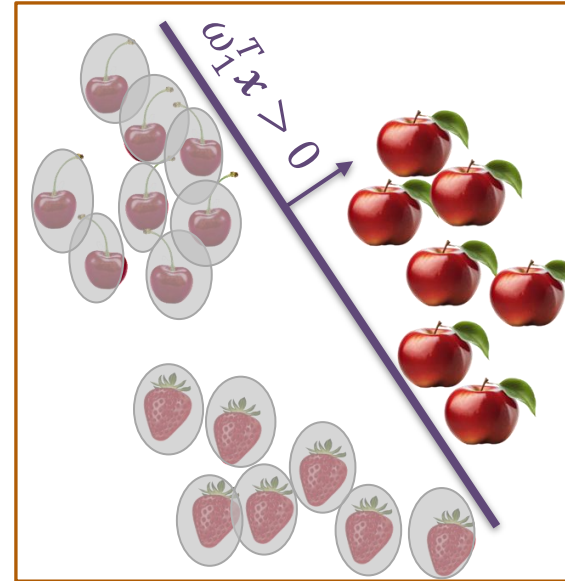
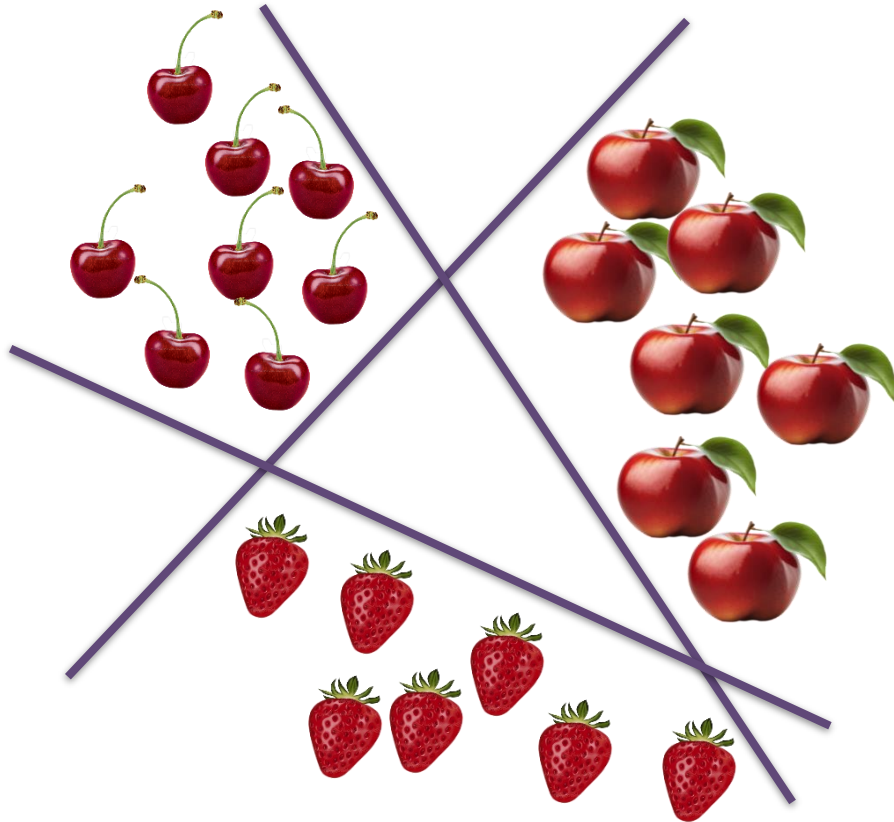
ONE-VS-REST



ONE-VS-REST



ONE-VS-REST



ONE-VS-ONE

Treinamos

$$\binom{K}{2} = \frac{K(K-1)}{2}$$

modelos de classificação binária separados:

$$h_{\omega}^{(kj)}(\mathbf{x}), k < j, k \in \{1, 2, \dots, K\}$$

$$\text{Para } k > j, h_{\omega}^{(jk)}(\mathbf{x}) = \mathbf{1} - h_{\omega}^{(kj)}(\mathbf{x}).$$

Para qualquer novo exemplo \mathbf{x} que desejamos prever escolhemos a classe com a pontuação máxima:

$$\hat{y} = \arg \max_{k \in \{1, 2, \dots, K\}} \sum_{j=1}^K h_{\omega}^{(kj)}(\mathbf{x}_{y=k,j})$$

$h_{\omega}^{(1j)}$ é um classificador binário projetado para reconhecer objetos da classe 1 entre objetos das classes 1 e j .

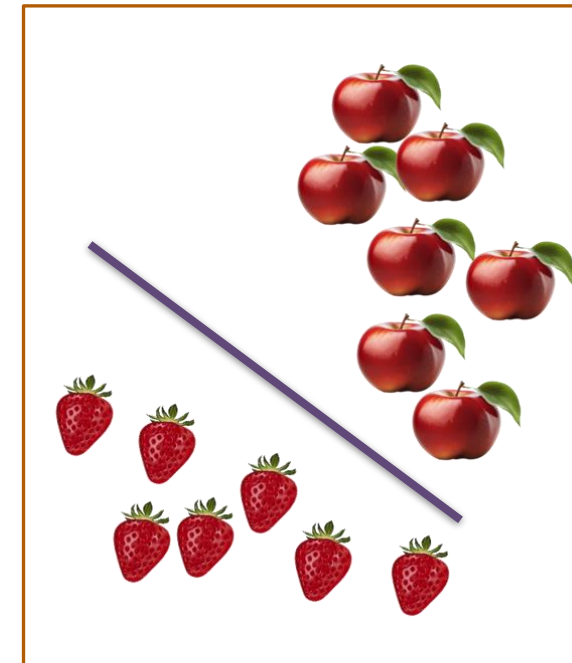
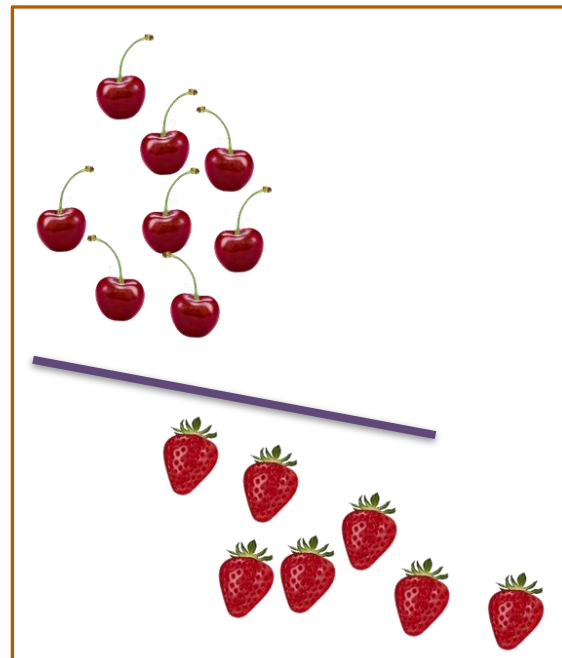
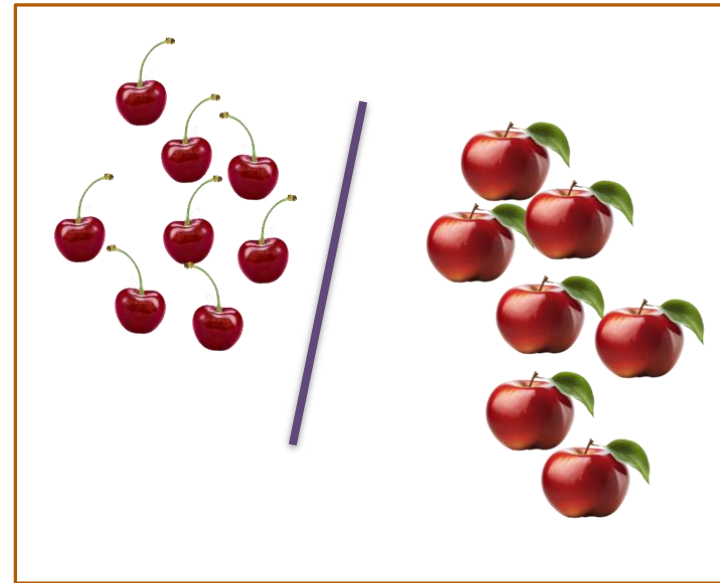
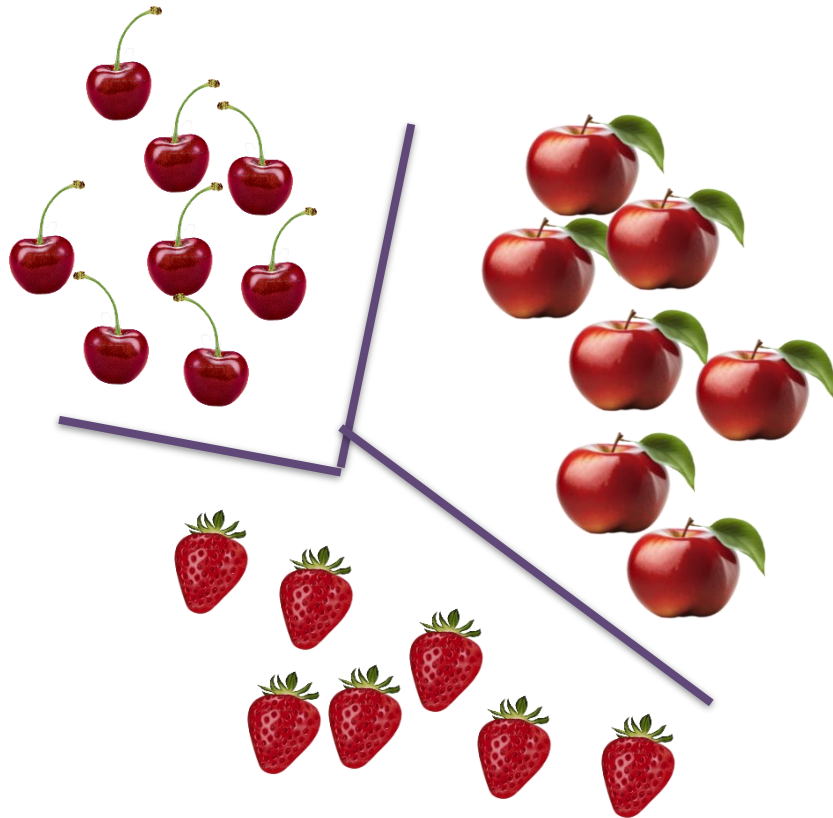
$$h_{\omega}^{(1j)}(\mathbf{x}_{y=1,j}) = p(y = 1 | \mathbf{x}_{y=1,j}; \omega)$$

$$h_{\omega}^{(2j)}(\mathbf{x}_{y=2,j}) = p(y = 2 | \mathbf{x}_{y=2,j}; \omega)$$

⋮

$$h_{\omega}^{(Kj)}(\mathbf{x}_{y=K,j}) = p(y = K | \mathbf{x}_{y=K,j}; \omega)$$

ONE-VS-ONE



MÉTODO MULTINOMIAL

Existe uma maneira de deixar o problema inerentemente multiclasse, com métodos que estimam as probabilidades condicionais $P(y = k | \mathbf{x}, \boldsymbol{\omega}), k \in \{1, 2, \dots, K\}$ de uma vez.

A generalização da regressão logística para múltiplas classes é conhecida como *regressão logística multinomial*.

Para estimar as probabilidades condicionais usamos a função softmax ao invés da sigmóide:

$$h_{\omega^{(k)}}(\mathbf{x}) = \frac{e^{\omega^{(k)} \cdot \mathbf{x}}}{\sum_{i=1}^K e^{\omega^{(i)} \cdot \mathbf{x}}}$$

$\omega^{(k)} \cdot \mathbf{x}$	$e^{\omega^{(k)} \cdot \mathbf{x}}$	$h_{\omega^{(k)}}(\mathbf{x})$
0	1	4,54E-05
0,2	1,221	5,54E-05
0,5	1,649	7,48E-05
1	2,718	0,000123
10	22026,47	0,999700
SOMA:	22033,05	1

A função softmax é legal porque garante que a soma de todas as nossas probabilidades de saída será igual a um.




ONE-HOT ENCODING

$$\mathbf{y}^{(i)} = \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ \vdots \\ y_K^{(i)} \end{bmatrix}$$

Pode receber valor 0 ou 1

$$\mathbf{h} = \begin{bmatrix} h^{(1)} \\ h^{(2)} \\ h^{(3)} \end{bmatrix} \begin{bmatrix} p(y = \text{apple} | \mathbf{x}; \omega) \\ p(y = \text{cherry} | \mathbf{x}; \omega) \\ p(y = \text{strawberry} | \mathbf{x}; \omega) \end{bmatrix} \sum = 1$$

$$\mathbf{y}^{(i)} = \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ y_3^{(i)} \end{bmatrix}$$

$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ 
 $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ 
 $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ 

REGRESSÃO MULTINOMIAL E FUNÇÃO CUSTO

$$h_{\omega^{(k)}}(\mathbf{x}) = \frac{e^{\omega^{(k)} \cdot \mathbf{x}}}{\sum_{i=1}^K e^{\omega^{(i)} \cdot \mathbf{x}}}$$



$$h_{\omega}(\mathbf{x}) = \begin{bmatrix} P(y=1|\mathbf{x}; \omega) \\ P(y=2|\mathbf{x}; \omega) \\ \vdots \\ P(y=K|\mathbf{x}; \omega) \end{bmatrix} = \frac{1}{\sum_{j=1}^K \exp(\omega^{(j)\top} \mathbf{x})} \begin{bmatrix} \exp(\omega^{(1)\top} \mathbf{x}) \\ \exp(\omega^{(2)\top} \mathbf{x}) \\ \vdots \\ \exp(\omega^{(K)\top} \mathbf{x}) \end{bmatrix}$$

$$J(\omega) = - \left[\sum_{i=1}^m \sum_{k=1}^K \boxed{1\{y^{(i)} = k\}} \log \frac{\exp(\omega^{(k)\top} \mathbf{x}^{(i)})}{\sum_{j=1}^K \exp(\omega^{(j)\top} \mathbf{x}^{(i)})} \right]$$

↓
 $1\{\blacksquare\}$ Função indicativa (1, se \blacksquare True; 0, cc)

QUAL É MELHOR?

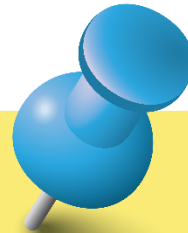
O método OvO é geralmente mais lento que OvR, devido à sua complexidade $\mathcal{O}(n^2)$ (são $K(K - 1)/2$ modelos). Em outras palavras, o método OvR é computacionalmente eficiente.

No entanto, o método OvO fornecerá maior precisão do que o método um contra o resto devido à comparação pareada de $K(K - 1)/2$ números de modelos.

Os modelos criados usando a estratégia OvO são facilmente interpretáveis.

OvR é o método padrão para classificação multiclasse.

No método multinomial, as probabilidades previstas são mais confiáveis do que os outros dois métodos.



Lição de casa

Você tem até domingo,
23:59hs. Estude os slides da
aula de hoje e entre no Moodle e
faça o teste com 3 questões
múltipla escolha.

NEVER GIVE UP



ACABOU...

Reveja a aula antes
de resolver os
exercícios.