APRENDIZADO DE MÁQUINA

FUNÇÃO DE PERDA (II)

Zenilton K. G. Patrocínio Jr
zenilton@pucminas.br

Outro Exemplo – Regressão Linear Multivariada

Neste tipo de modelo, x e y são vetores de dimensões m e k, isto é, $x \in \mathbb{R}^m$ e $y \in \mathbb{R}^k$, sendo o modelo dado por

$$y = Ax$$

em que A é uma matriz de coeficientes (ou parâmetros) de dimensões $k \times m$.

Outro Exemplo – Regressão Linear Multivariada

Neste tipo de modelo, x e y são vetores de dimensões m e k, isto é, $x \in \mathbb{R}^m$ e $y \in \mathbb{R}^k$, sendo o modelo dado por

$$y = Ax$$

em que A é uma matriz de coeficientes (ou parâmetros) de dimensões $k \times m$.

Assim, como antes, o risco empírico é dado pela soma da perda em todos os n pares, ou ainda,

$$L = \sum_{i=1}^{n} (Ax_i - y_i)^2 = \sum_{i=1}^{n} \left(x_i^T A^T - y_i^T\right)(Ax_i - y_i)$$

Outro Exemplo – Regressão Linear Multivariada

Neste caso, para se minimizar o valor do risco empírico é necessário se igualar a zero o gradiente de L em relação à matriz A, isto é

$$\nabla_A L = 0$$

considerando, assim, que L é uma função da matriz A

Outro Exemplo – Regressão Linear Multivariada

Neste caso, para se minimizar o valor do risco empírico é necessário se igualar a zero o gradiente de L em relação à matriz A, isto é

$$\nabla_A L = 0$$

considerando, assim, que L é uma função da matriz A

O gradiente $\nabla_A L(A)$ representa o vetor de derivadas parciais

$$\nabla_A L(A) = \left[\frac{\partial L}{\partial A_{11}}, \frac{\partial L}{\partial A_{12}}, \cdots, \frac{\partial L}{\partial A_{21}}, \frac{\partial L}{\partial A_{22}}, \cdots \right]^T$$

Outro Exemplo – Regressão Linear Multivariada

Neste caso, para se minimizar o valor do risco empírico é necessário se igualar a zero o gradiente de L em relação à matriz A, isto é

$$\nabla_A L = 0$$

considerando, assim, que L é uma função da matriz A

O gradiente $\nabla_A L(A)$ representa o vetor de derivadas parciais

$$\nabla_A L(A) = \left[\frac{\partial L}{\partial A_{11}}, \frac{\partial L}{\partial A_{12}}, \cdots, \frac{\partial L}{\partial A_{21}}, \frac{\partial L}{\partial A_{22}}, \cdots \right]^T$$

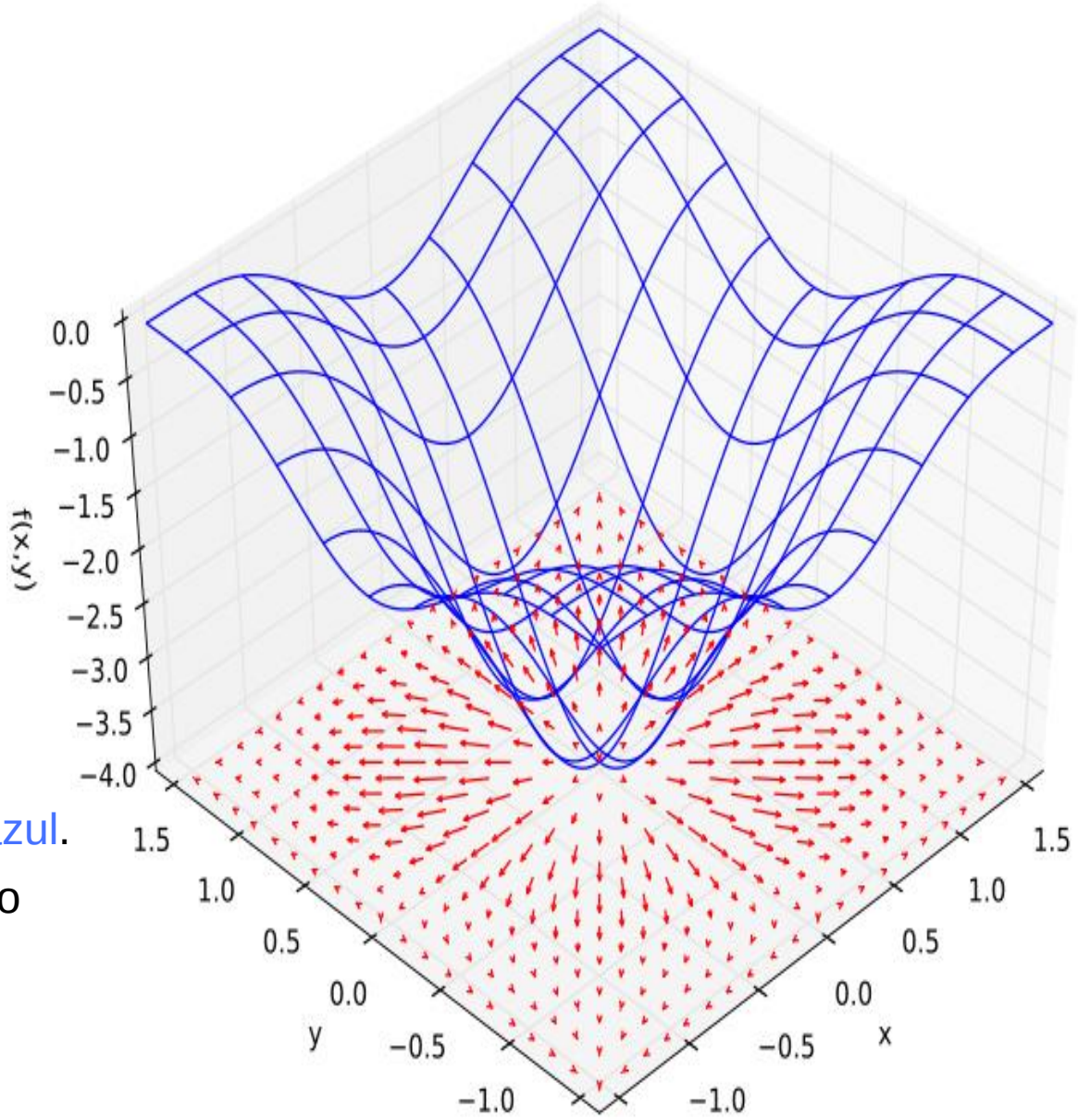
em que, por exemplo, $\frac{\partial L}{\partial A_{11}}$ mede o quão rápido varia a perda L em relação a uma variação do coeficiente A_{11} da matriz A

Outro Exemplo – Regressão Linear Multivariada

Quando $\nabla_A L(A) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção

Outro Exemplo – Regressão Linear Multivariada

Quando $\nabla_A L(A) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção



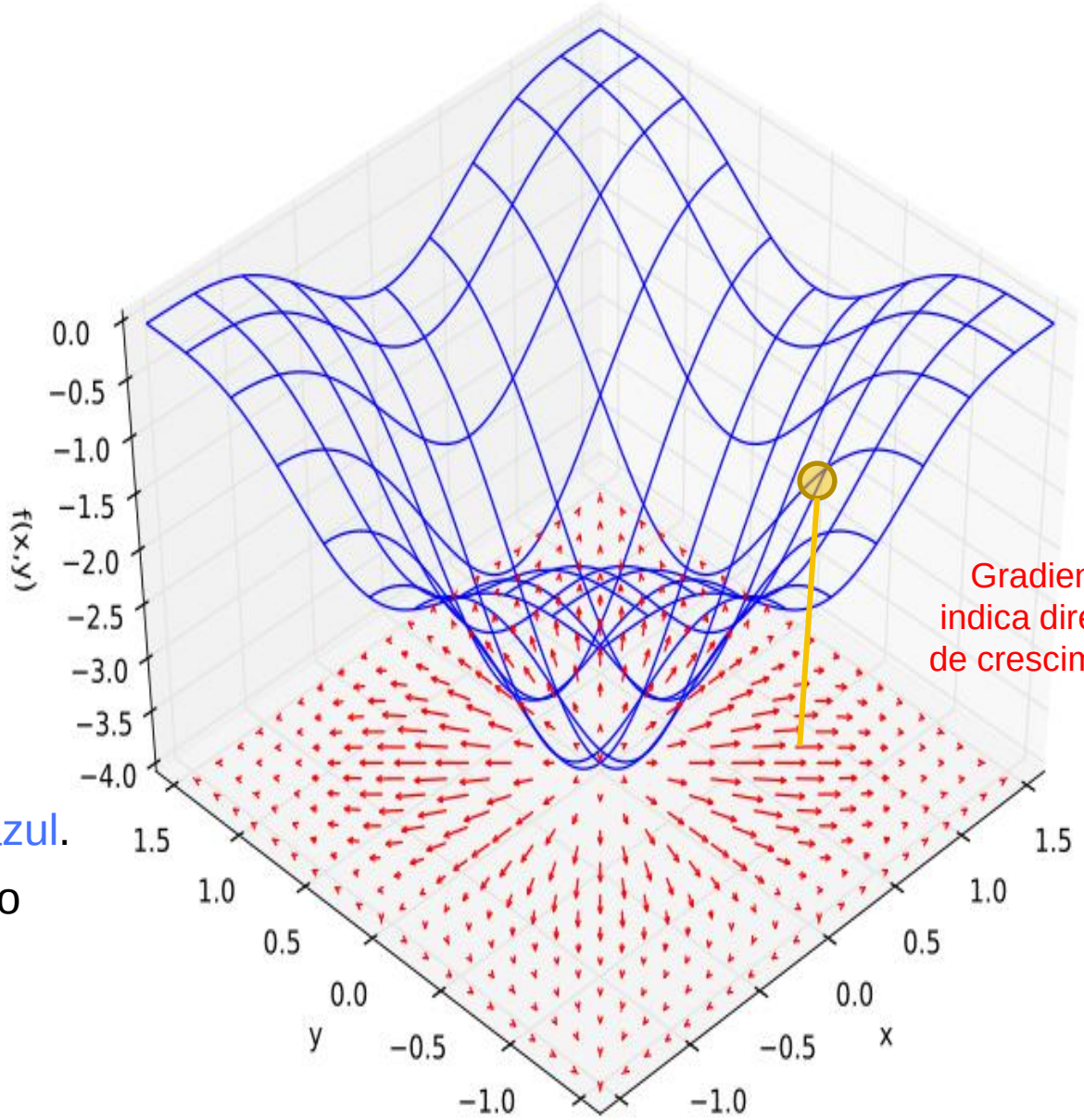
A superfície da perda aparece em azul.

Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Outro Exemplo – Regressão Linear Multivariada

Quando $\nabla_A L(A) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção



A superfície da perda aparece em azul.

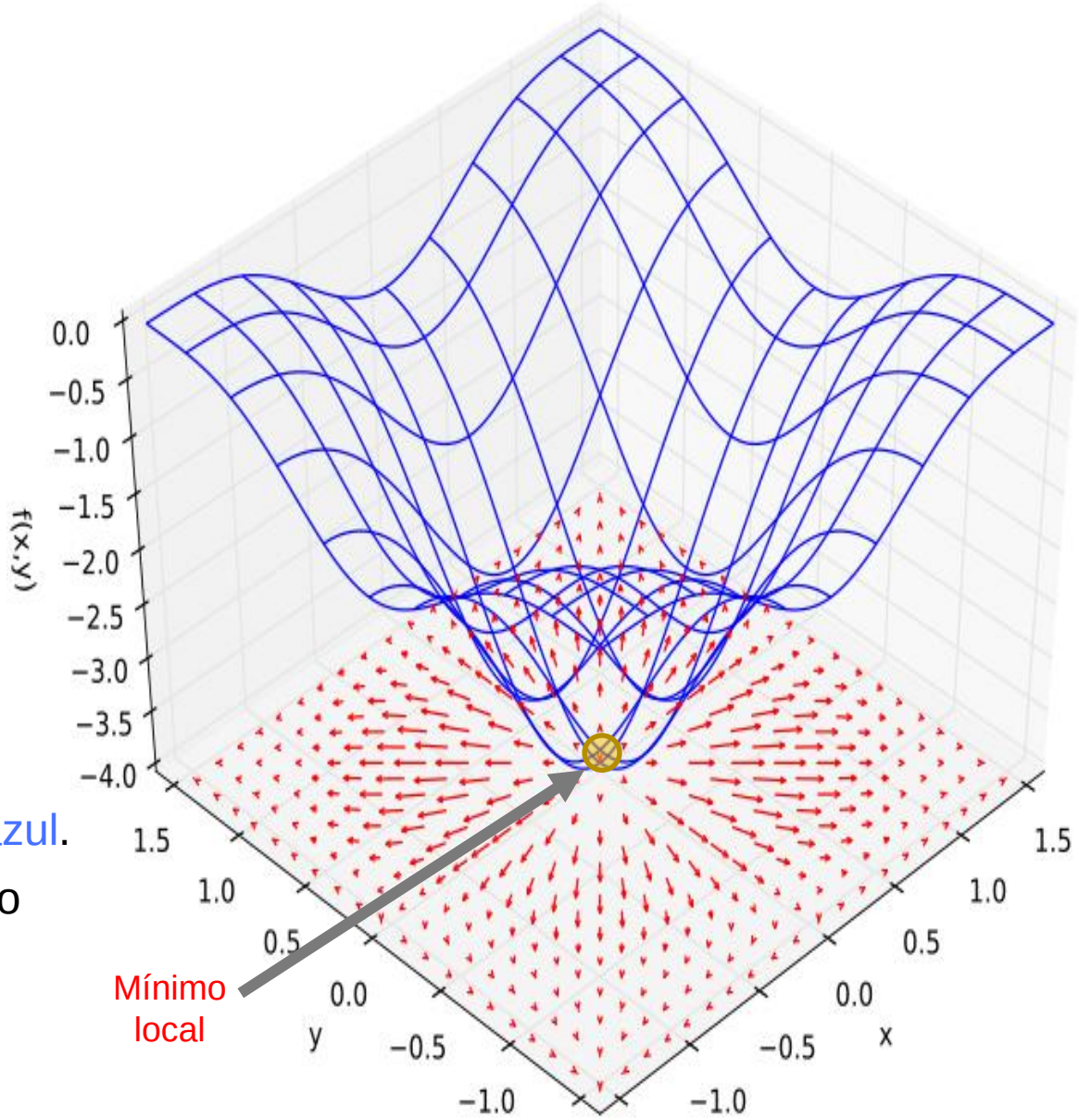
Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Outro Exemplo – Regressão Linear Multivariada

Quando $\nabla_A L(A) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção

Portanto, obteve-se um ótimo local



A superfície da perda aparece em azul.

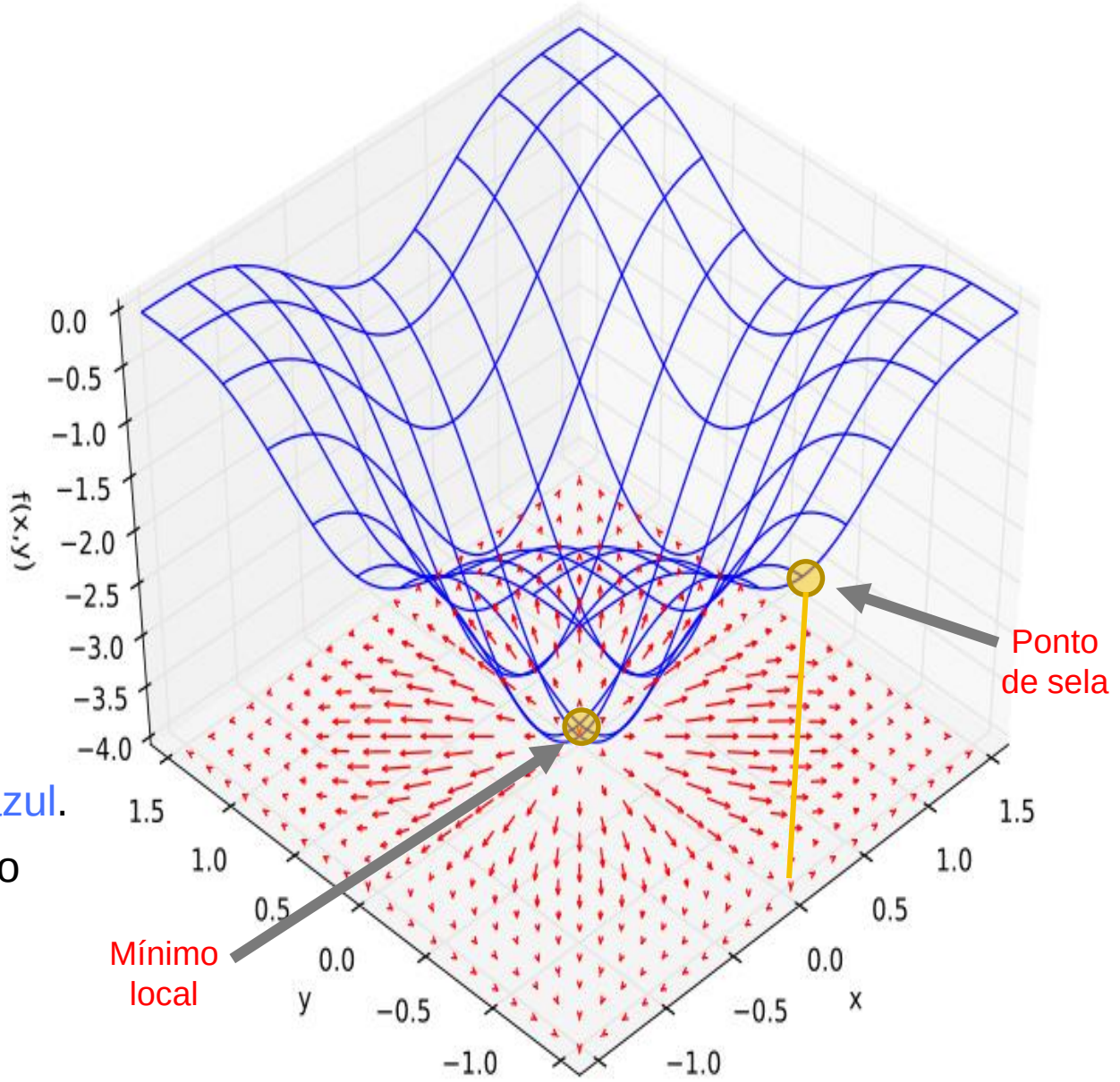
Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Outro Exemplo – Regressão Linear Multivariada

Quando $\nabla_A L(A) = 0$, então todas as derivadas parciais são nulas, ou ainda, a perda não se modifica em nenhuma direção

Portanto, obteve-se um ótimo local (ou, pelo menos, um ponto de sela)



A superfície da perda aparece em azul.

Os gradientes são mostrados abaixo como setas vermelhas.

O gradiente é zero no mínimo.

Função de Perda de Entropia Cruzada

Suponha que a predição $\hat{y} = f(x)$ seja a ***probabilidade de x ser rotulado na classe alvo*** em um problema com duas classes

Função de Perda de Entropia Cruzada

Suponha que a predição $\hat{y} = f(x)$ seja a ***probabilidade de x ser rotulado na classe alvo*** em um problema com duas classes

Sob essa suposição, pode-se calcular a ***probabilidade da classificação correta*** e maximizá-la

Função de Perda de Entropia Cruzada

Suponha que a predição $\hat{y} = f(x)$ seja a ***probabilidade de x ser rotulado na classe alvo*** em um problema com duas classes

Sob essa suposição, pode-se calcular a ***probabilidade da classificação correta*** e maximizá-la

A probabilidade da classificação correta para uma entrada é

$$p_{correta} = \begin{cases} \hat{y} & \text{se} \quad y = 1 \quad \text{(probabilidade de verdadeito positivo)} \\ 1 - \hat{y} & \text{se} \quad y = 0 \quad \text{(probabilidade de verdadeiro negativo)} \end{cases}$$

Função de Perda de Entropia Cruzada

Suponha que a predição $\hat{y} = f(x)$ seja a ***probabilidade de x ser rotulado na classe alvo*** em um problema com duas classes

Sob essa suposição, pode-se calcular a ***probabilidade da classificação correta*** e maximizá-la

A probabilidade da classificação correta para uma entrada é

$$p_{correta} = \begin{cases} \hat{y} & \text{se} \quad y = 1 \quad (\text{probabilidade de verdadeito positivo}) \\ 1 - \hat{y} & \text{se} \quad y = 0 \quad (\text{probabilidade de verdadeiro negativo}) \end{cases}$$

Podendo ser representada por uma expressão simples da seguinte forma

$$p_{correta} = y\,\hat{y} + \ (1 - y)(1 - \hat{y})$$

Função de Perda de Entropia Cruzada

Maximizar a ***probabilidade da classificação correta*** equivale a ***minimizar***

- $p_{correta}$ como a perda de cada observação, isto é

$$-p_{correta} = -y\,\hat{y} - (1 - y)(1 - \hat{y})$$

Função de Perda de Entropia Cruzada

Maximizar a ***probabilidade da classificação correta*** equivale a ***minimizar*** $-p_{correta}$ como a perda de cada observação, isto é

$$-p_{correta} = -y\,\hat{y} - (1 - y)(1 - \hat{y})$$

A seguir, somam-se as perdas de todas as observações, de forma que ao minimiza-la obtém-se um modelo de menor risco empírico

Função de Perda de Entropia Cruzada

Maximizar a ***probabilidade da classificação correta*** equivale a ***minimizar*** $-p_{correta}$ como a perda de cada observação, isto é

$$-p_{correta} = -y\,\hat{y} - (1 - y)(1 - \hat{y})$$

A seguir, somam-se as perdas de todas as observações, de forma que ao minimiza-la obtém-se um modelo de menor risco empírico

Mas é muito mais conveniente se usar o ***log negativo da probabilidade de um resultado correto***

$$-\log p_{correta} = -y \log \hat{y} - (1 - y)\log(1 - \hat{y})$$

Função de Perda de Entropia Cruzada

A perda de entropia cruzada é ***a probabilidade logarítmica negativa de que todos os rótulos estejam corretos*** – como os erros dos rótulos são independentes, devemos multiplicá-los para obter a probabilidade total

Função de Perda de Entropia Cruzada

A perda de entropia cruzada é ***a probabilidade logarítmica negativa de que todos os rótulos estejam corretos*** – como os erros dos rótulos são independentes, devemos multiplicá-los para obter a probabilidade total

O uso de ***logs*** transforma este produto em uma soma. Dessa forma, é obtida a ***Perda de Entropia Cruzada*** (ou "***cross-entropy loss***")

$$L = -\sum_{i=1}^{n} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$$

Função de Perda de Entropia Cruzada

A perda de entropia cruzada é ***a probabilidade logarítmica negativa de que todos os rótulos estejam corretos*** – como os erros dos rótulos são independentes, devemos multiplicá-los para obter a probabilidade total

O uso de ***logs*** transforma este produto em uma soma. Dessa forma, é obtida a ***Perda de Entropia Cruzada*** (ou "***cross-entropy loss***")

$$L = -\sum_{i=1}^{n} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$$

Genericamente, a ***Perda de Entropia Cruzada*** compara uma distribuição alvo y_i (que pode não ser binária) com uma distribuição $\hat{y}$ dada pelo modelo