APRENDIZADO DE MÁQUINA (III)

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

Função de Perda

Pode não haver um valor alvo y “verdadeiro” para uma observação x, ou seja, pode haver vários “diferentes” valores de y para o mesmo x

Função de Perda

Pode não haver um valor alvo y “verdadeiro” para uma observação x, ou seja, pode haver vários “diferentes” valores de y para o mesmo x

Também pode haver ruído ou efeitos não modelados no conjunto de dados; portanto, mesmo se houver um único y para um dado x, pode ser impossível predizê-lo com exatidão

Função de Perda

Pode não haver um valor alvo y "verdadeiro" para uma observação x, ou seja, pode haver vários "diferentes" valores de y para o mesmo x

Também pode haver ruído ou efeitos não modelados no conjunto de dados; portanto, mesmo se houver um único y para um dado x, pode ser impossível predizê-lo com exatidão

Em vez disso, tenta-se predizer um valor "próximo" do valor alvo observado

Função de Perda

Neste caso, usa-se durante o treinamento uma **função de perda** para medir a proximidade de tais predições feitas em relação aquelas do conjunto de dados que foram previamente observadas

Função de Perda

Neste caso, usa-se durante o treinamento uma **função de perda** para medir a proximidade de tais predições feitas em relação aquelas do conjunto de dados que foram previamente observadas

Uma **função de perda** mede a diferença entre uma predição do valor alvo e o valor disponível no conjunto de treinamento

Função de Perda

Neste caso, usa-se durante o treinamento uma **função de perda** para medir a proximidade de tais predições feitas em relação aquelas do conjunto de dados que foram previamente observadas

Uma **função de perda** mede a diferença entre uma predição do valor alvo e o valor disponível no conjunto de treinamento

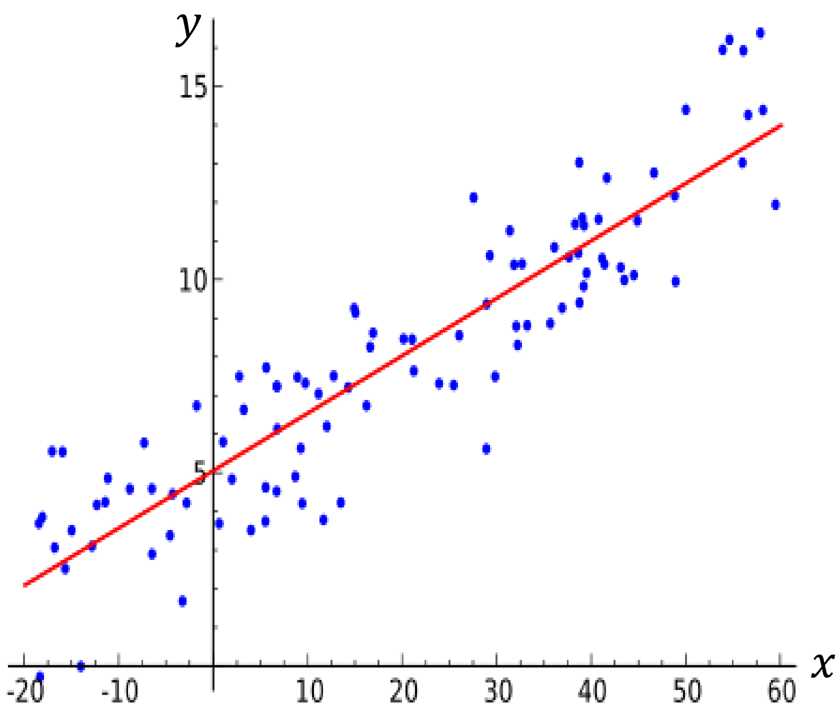
Exemplo: Perda quadrática (ou "***squared loss***") $L_2(\hat{y}, y) = (\hat{y} - y)^2$ em que $\hat{y} = f(x)$ representa a predição feita pelo modelo para o par de dados (x, y)

Exemplo – Regressão Linear

Nesse tipo de modelo bem simples, $\hat{y} = f(x) = ax + b$ sendo x um valor real, enquanto a e b são constantes reais

Exemplo – Regressão Linear

Nesse tipo de modelo bem simples, $\hat{y} = f(x) = ax + b$ sendo x um valor real, enquanto a e b são constantes reais



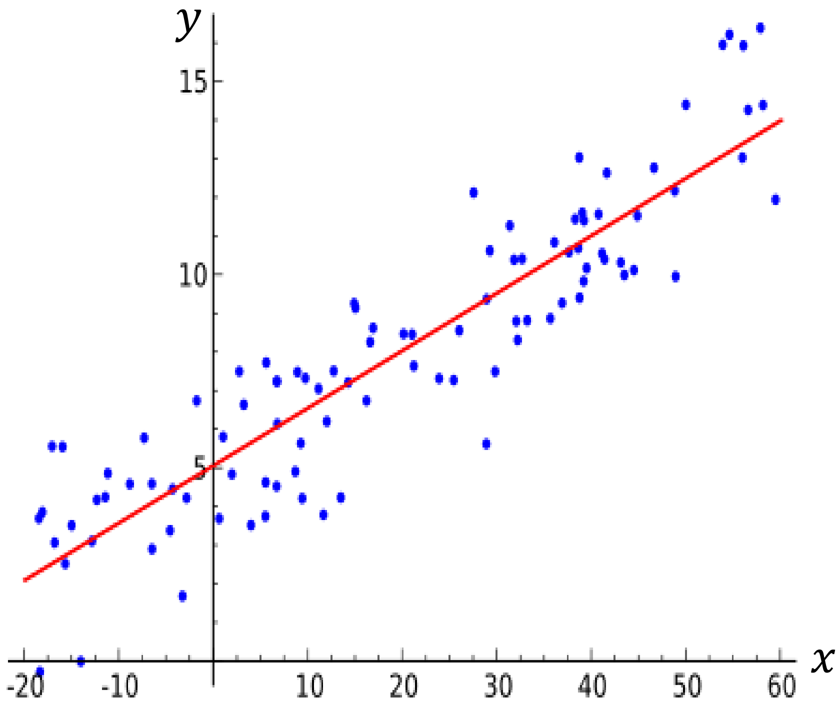
Os n pares de dados (x, y) são os pontos azuis

Já o modelo é representado pela linha vermelha

Exemplo – Regressão Linear

Nesse tipo de modelo bem simples, $\hat{y} = f(x) = ax + b$ sendo x um valor real, enquanto a e b são constantes reais

A perda quadrática $L_2(\hat{y}, y) = (\hat{y} - y)^2$ avalia a diferença (ou distância) entre predição $\hat{y}$ e dado y



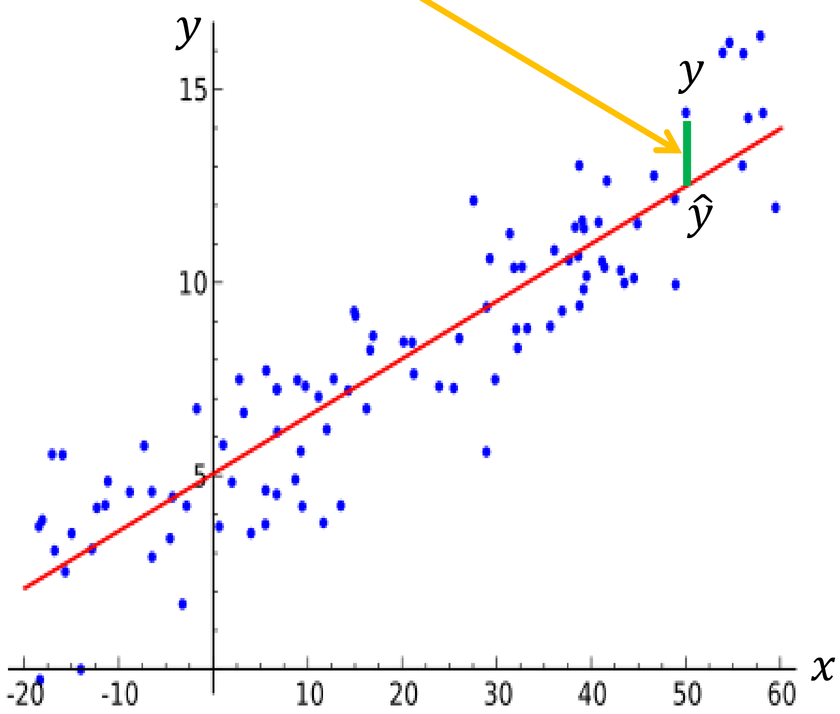
Os n pares de dados (x, y) são os pontos azuis

Já o modelo é representado pela linha vermelha

Exemplo – Regressão Linear

Nesse tipo de modelo bem simples, $\hat{y} = f(x) = ax + b$ sendo x um valor real, enquanto a e b são constantes reais

A perda quadrática $L_2(\hat{y}, y) \ = \ (\hat{y} - y)^2$ avalia a diferença (ou distância) entre predição $\hat{y}$ e dado y



Os n pares de dados (x, y) são os pontos azuis

Já o modelo é representado pela linha vermelha

Exemplo – Regressão Linear

A perda total pode ser obtida somando-se a perda em todos os n pares

$$L = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Exemplo – Regressão Linear

A perda total pode ser obtida somando-se a perda em todos os n pares

$$L = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Uma vez que a predição $\hat{y}_i$ para um x_i qualquer é dada por $ax_i + b$, pode-se rescrever a perda total como

$$L = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

Exemplo – Regressão Linear

A perda total pode ser obtida somando-se a perda em todos os n pares

$$L = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Uma vez que a predição $\hat{y}_i$ para um x_i qualquer é dada por $ax_i + b$, pode-se rescrever a perda total como

$$L = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

Para se encontrar os valores ótimos de a e b, pode-se aplicar o cálculo diferencial, igualando à zero as derivadas da perda total em relação a e b, isto é

$$\frac{\partial L}{\partial a} = 0 \quad \text{e} \quad \frac{\partial L}{\partial b} = 0$$

Exemplo – Regressão Linear

Dessa forma, para se minimizar a perda e obter os valores ótimos de a e b, deve-se fazer

$$\frac{\partial L}{\partial a} = 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i = 0$$

$$\frac{\partial L}{\partial b} = 2a\sum_{i=1}^{n} x_i \quad + 2bn - 2\sum_{i=1}^{n} y_i \qquad = 0$$

Exemplo – Regressão Linear

Dessa forma, para se minimizar a perda e obter os valores ótimos de a e b, deve-se fazer

$$\frac{dL}{da} = 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i = 0$$

$$\frac{dL}{db} = 2a\sum_{i=1}^{n} x_i + 2bn - 2\sum_{i=1}^{n} y_i = 0$$

Como os somatórios nessas equações são constantes para os n pares de dados, tem-se duas equações **lineares** em a e b, que podem ser facilmente resolvidas

Exemplo – Regressão Linear

Dessa forma, para se minimizar a perda e obter os valores ótimos de a e b, deve-se fazer

$$\frac{dL}{da} = 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i = 0$$

$$\frac{dL}{db} = 2a\sum_{i=1}^{n} x_i + 2bn - 2\sum_{i=1}^{n} y_i = 0$$

Como os somatórios nessas equações são constantes para os n pares de dados, tem-se duas equações **lineares** em a e b, que podem ser facilmente resolvidas obtendo-se valores $\boldsymbol{\tilde{a}}$ e $\boldsymbol{\tilde{b}}$

O modelo $\boldsymbol{f(x) = \tilde{a}x + \tilde{b}}$ com estes valores é o **modelo de perda mínima**

Minimização de Risco

Recapitulando... obteve-se o valor das constantes a e b que minimizam a perda quadrática sobre alguns dados que ***nós já possuíamos***

Minimização de Risco

Recapitulando... obteve-se o valor das constantes a e b que minimizam a perda quadrática sobre alguns dados que ***nós já possuíamos***

Porém o que se quer na verdade é predizer os valores de y para observações x que ***nós ainda não dispomos***, isto é, gostaríamos de minimizar a perda esperada sobre novos dados, ou ainda, $\mathbb{E}[(\hat{y} - y)^2]$

Minimização de Risco

Recapitulando... obteve-se o valor das constantes a e b que minimizam a perda quadrática sobre alguns dados que ***nós já possuíamos***

Porém o que se quer na verdade é predizer os valores de y para observações x que ***nós ainda não dispomos***, isto é, gostaríamos de minimizar a perda esperada sobre novos dados, ou ainda, $\mathbb{E}[(\hat{y} - y)^2]$

Tal perda esperada é denominada **risco**

Minimização de Risco

Na verdade, minimizou-se um valor de perda obtido sobre um número finito de dados disponíveis que é chamado de **risco empírico**

Minimização de Risco

Na verdade, minimizou-se um valor de perda obtido sobre um número finito de dados disponíveis que é chamado de **risco empírico**

Dessa forma, o aprendizado de máquina aproxima modelos que minimizam o risco por meio de modelos que minimizem o risco empírico

Minimização de Risco

Na verdade, minimizou-se um valor de perda obtido sobre um número finito de dados disponíveis que é chamado de **risco empírico**

Dessa forma, o aprendizado de máquina aproxima modelos que minimizam o risco por meio de modelos que minimizem o risco empírico

Geralmente, minimizar o risco empírico (perda de dados) em vez do risco real funciona bem, mas pode falhar se:

- A amostra de dados for enviesada, ou
- Não houver dados suficientes para estimar com precisão os parâmetros do modelo

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, (usando perda quadrática)

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

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

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

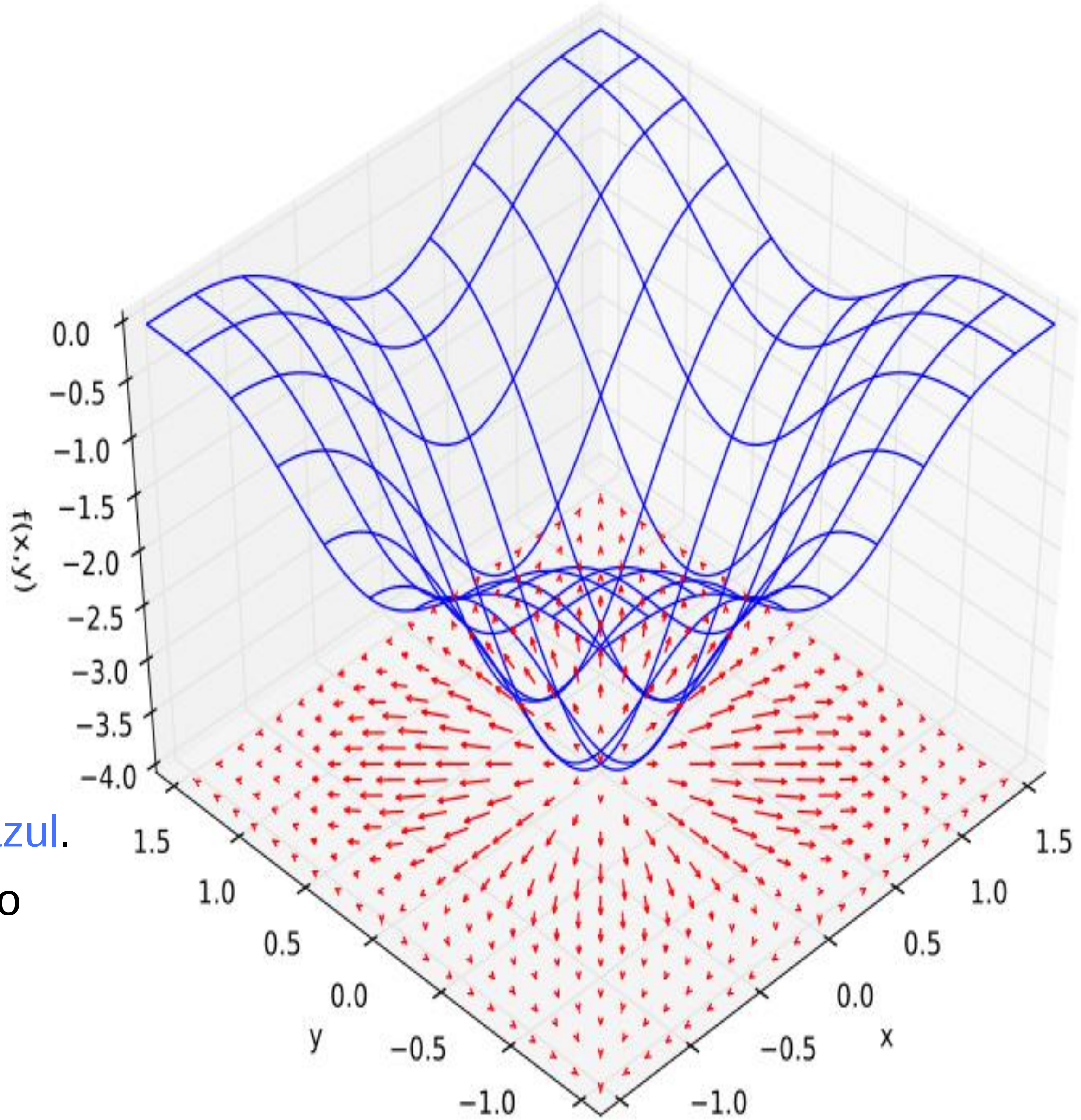
em que, por exemplo, $\frac{\partial L}{\partial A_{11}}$ mede o quão rápido varia a perda L em relação a uma variação do coeficiente A_{11} da matriz A

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



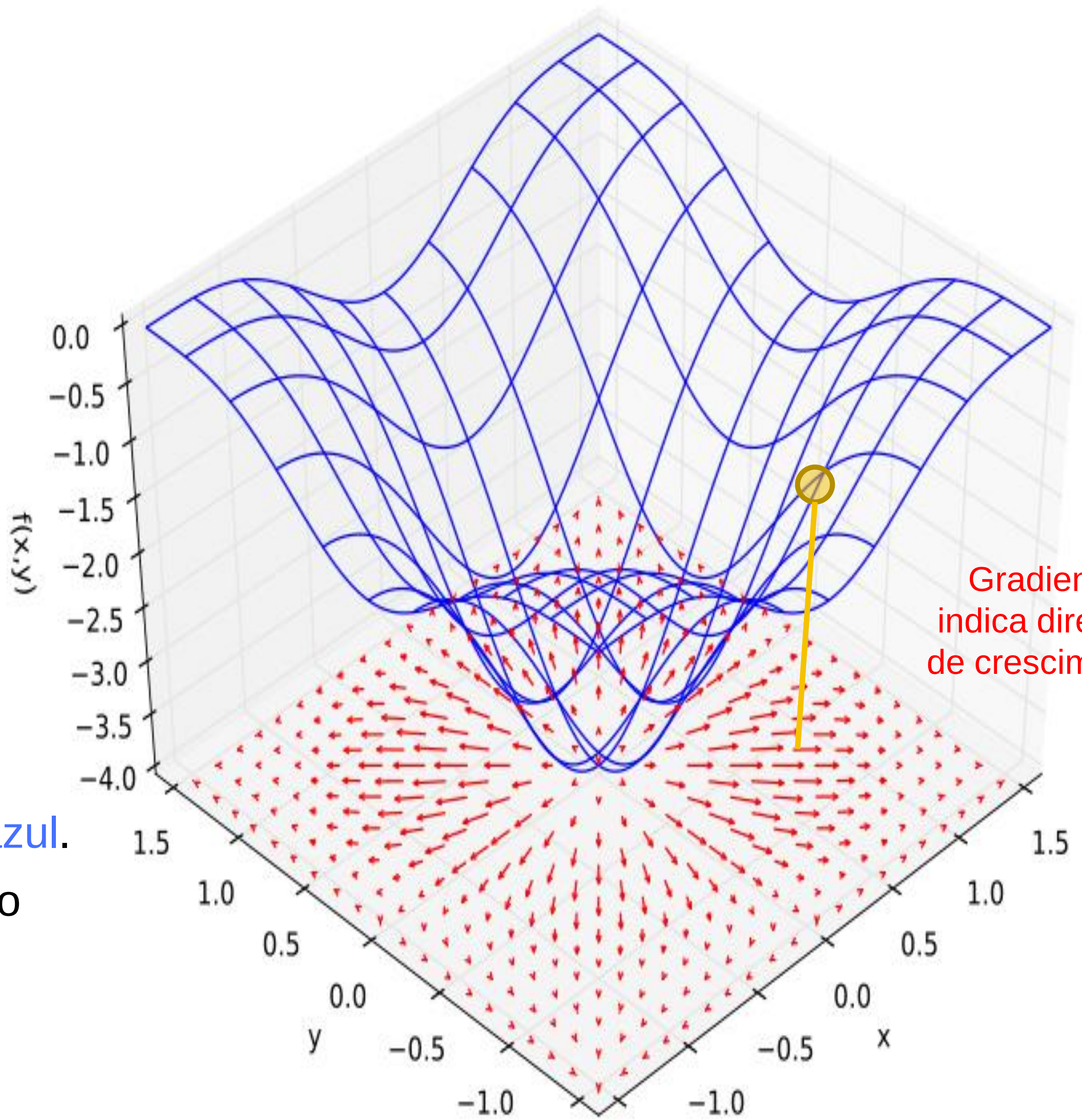
A superfície da perda aparece em azul.

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



Gradiente indica direção de crescimento

A superfície da perda aparece em azul.

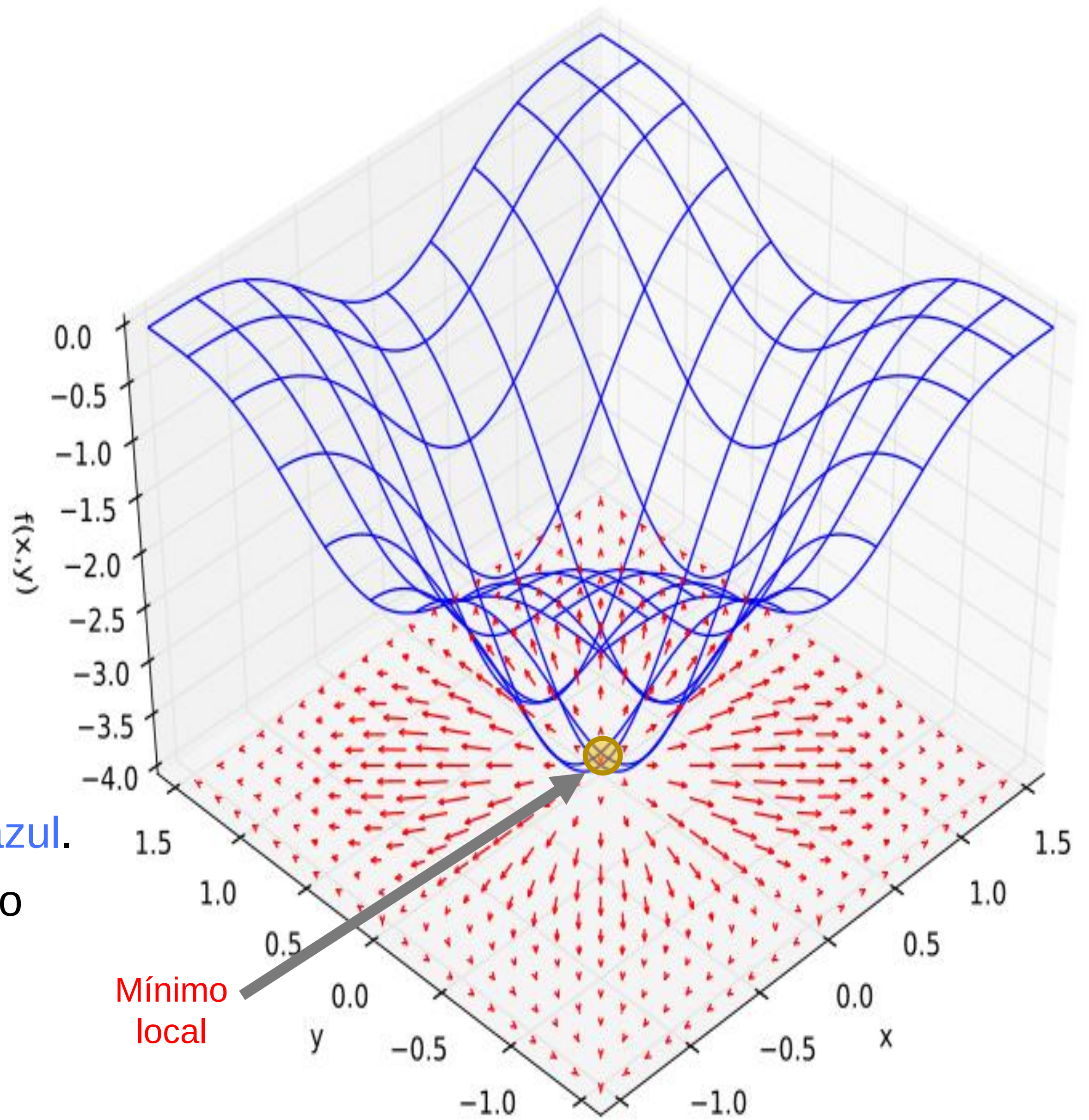
Os gradientes são mostrados abaixo como setas vermelhas.

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.

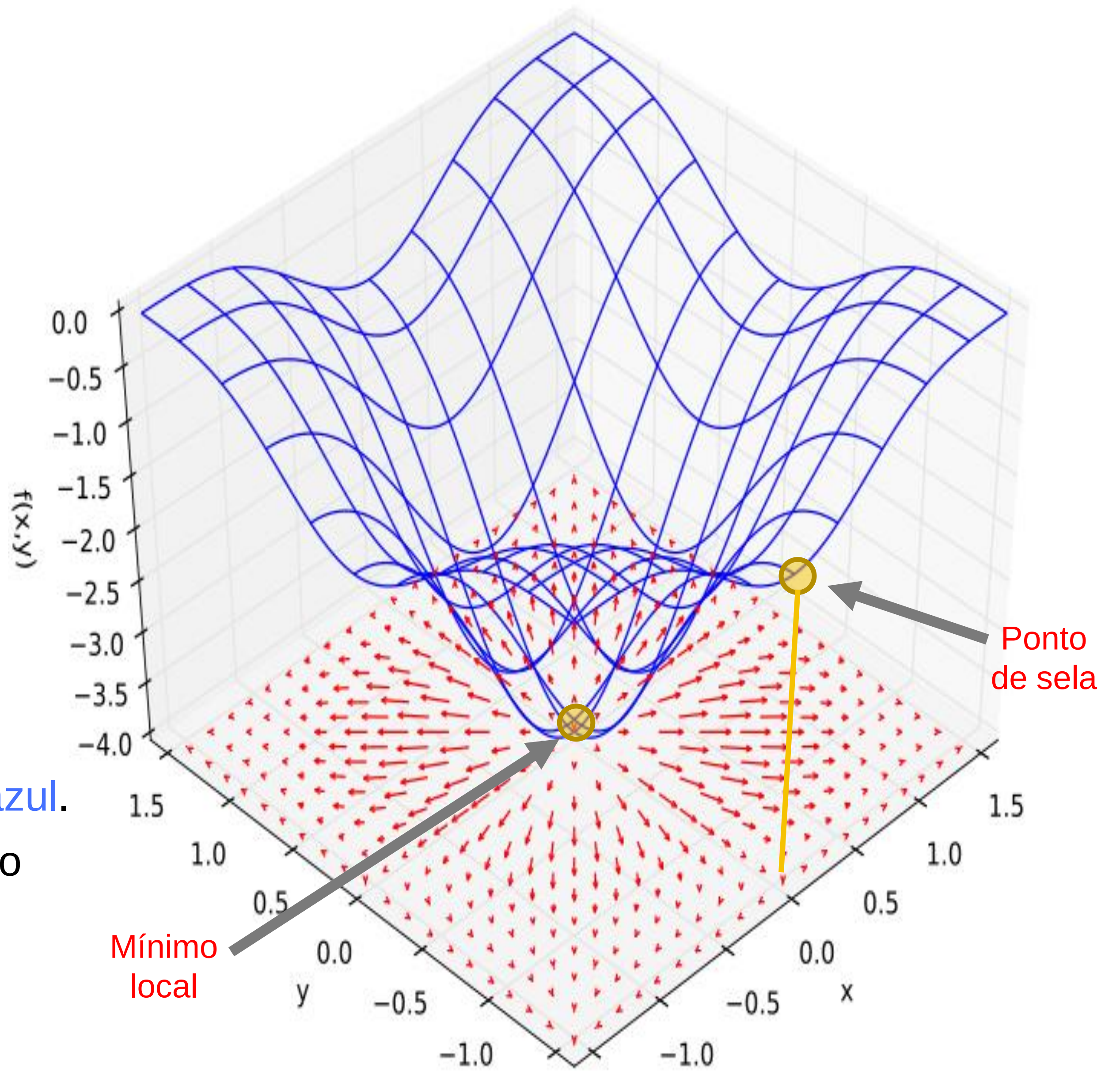
Os gradientes são mostrados abaixo como setas vermelhas.

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

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

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

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

Função de Perda de Articulação

A **função perda de articulação** (“***hinge loss***”) usa a noção “margem máxima” buscando obter fronteiras com a maior distância dos dados

Função de Perda de Articulação

A **função perda de articulação** (“***hinge loss***”) usa a noção “margem máxima” buscando obter fronteiras com a maior distância dos dados

A noção de ***margem*** pode ser representada pela diferença entre o “score” da classe correta e uma outra classe qualquer

Função de Perda de Articulação

A **função perda de articulação** (“***hinge loss***”) usa a noção “margem máxima” buscando obter fronteiras com a maior distância dos dados

A noção de ***margem*** pode ser representada pela diferença entre o “score” da classe correta e uma outra classe qualquer

Seja y a classe correta correspondendo a x e j uma outra classe qualquer. Então a perda para a classe j é dada por

$$\max\left(0, f_j(x) - f_y(x) + 1\right)$$

Função de Perda de Articulação

A **função perda de articulação** (“***hinge loss***”) usa a noção “margem máxima” buscando obter fronteiras com a maior distância dos dados

A noção de ***margem*** pode ser representada pela diferença entre o “score” da classe correta e uma outra classe qualquer

Seja y a classe correta correspondendo a x e j uma outra classe qualquer. Então a perda para a classe j é dada por

$$\max\left(0, f_j(x) - f_y(x) + 1\right)$$

Somando-se para todas as classes j que foram diferente de y, obtém-se

$$L = \sum_{j \neq y} \max\left(0, f_j(x) - f_y(x) + 1\right)$$

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$

são



gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**

Valores em negrito são “scores” para a classe correta.

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$ são

gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$

são

x_i →

				
y_i →	**gato**	**3,2**	1,3	2,2
	carro	5,1	**4,9**	2,5
	rã	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$ →

$\boldsymbol{y_i}$	→	**gato**	**3,2**	1,3	2,2
$\boldsymbol{j}$	→	**carro**	5,1	**4,9**	2,5
		rã	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$ →   

$\boldsymbol{y_i}$ →	**gato**	**3,2**	1,3	2,2
$\boldsymbol{j}$ →	**carro**	5,1	**4,9**	2,5
	rã	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 5,1 – 3,2 + 1)

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$

são

$\boldsymbol{x_i}$ →

$\boldsymbol{y_i}$ →	**gato**	**3,2**	1,3	2,2
$\boldsymbol{j}$ →	**carro**	5,1	**4,9**	2,5
	rã	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 5,1 – 3,2 + 1)

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$ →



$\boldsymbol{y_i}$ →	**gato**	**3,2**	1,3	2,2
	carro	5,1	**4,9**	2,5
$\boldsymbol{j}$ →	**rã**	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max\left(0, s_j - s_{y_i} + 1\right)$$

= max(0, 5,1 – 3,2 + 1) +
max(0, -1,7 – 3,2 + 1)

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$

$\boldsymbol{y_i}$	**gato**	**3,2**	1,3	2,2
	carro	5,1	**4,9**	2,5
$\boldsymbol{j}$	**rã**	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 5,1 – 3,2 + 1) + max(0, -1,7 – 3,2 + 1)

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$ →



$\boldsymbol{y_i}$ → **gato**	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 5,1 – 3,2 + 1) +
max(0, -1,7 – 3,2 + 1)

= max(0, 2,9) + max(0, -3,9)

= 2,9 + 0

= 2,9

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = \; f(x, W) = Wx$ são

$\boldsymbol{x_i}$



$\boldsymbol{y_i}$ **gato**	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**
Perda:	**2,9**		

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max\left(0, s_j - s_{y_i} + 1\right)$$

= max(0, 5,1 – 3,2 + 1) +
max(0, -1,7 – 3,2 + 1)

= max(0, 2,9) + max(0, -3,9)

= 2,9 + 0

= 2,9

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$

	gato	**3,2**	1,3	2,2
$\boldsymbol{y_i} \rightarrow$	**carro**	5,1	**4,9**	2,5
	rã	-1,7	2,0	**-3,1**
	Perda:	**2,9**		

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$

	gato	**3,2**	1,3	2,2
$\boldsymbol{y_i}$	**carro**	5,1	**4,9**	2,5
	rã	-1,7	2,0	**-3,1**
	Perda:	**2,9**		

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 1,3 – 4,9 + 1) +
max(0, 2,0 – 4,9 + 1)

= max(0, -2,6) + max(0, -1,9)

= 0 + 0

= 0

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$

	gato	**3,2**	1,3	2,2
$\boldsymbol{y_i}$	**carro**	5,1	**4,9**	2,5
	rã	-1,7	2,0	**-3,1**
	Perda:	**2,9**	**0**	

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 1,3 – 4,9 + 1) +
max(0, 2,0 – 4,9 + 1)

= max(0, -2,6) + max(0, -1,9)

= 0 + 0

= 0

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$

	gato	**3,2**	1,3	2,2
	carro	5,1	**4,9**	2,5
$\boldsymbol{y_i}$ →	**rã**	-1,7	2,0	**-3,1**
	Perda:	**2,9**	**0**	

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$

são

$\boldsymbol{x_i}$



	gato	**3,2**	1,3	2,2
	carro	5,1	**4,9**	2,5
$\boldsymbol{y_i} \rightarrow$	**rã**	-1,7	2,0	**-3,1**
	Perda:	**2,9**	**0**	

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro)

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

= max(0, 2,2 – (-3,1) + 1) +
max(0, 2,5 – (-3,1) + 1)

= max(0, 6,3) + max(0, 6,6)

= 6,3 + 6,6

= 12,9

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$ são

$\boldsymbol{x_i}$



	gato	**3,2**	1,3	2,2
	carro	5,1	**4,9**	2,5
$\boldsymbol{y_i} \rightarrow$	**rã**	-1,7	2,0	**-3,1**
	Perda:	**2,9**	**0**	**12,9**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro),

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max\left(0, s_j - s_{y_i} + 1\right)$$

= max(0, 2,2 – (-3,1) + 1) + max(0, 2,5 – (-3,1) + 1)

= max(0, 6,3) + max(0, 6,6)

= 6,3 + 6,6

= 12,9

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$

são



gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**
Perda:	**2,9**	**0**	**12,9**

Dada uma amostra (x_i, y_i) em que x_i é a imagem e y_i é o rótulo da classe (um valor inteiro),

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Assim a perda total é dada pela a soma de todas as perdas

$$L = \sum_{i=1}^{N} L_i$$

L = 2,9 + 0 + 12,9 = **15,8**

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$ são

			
gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**
Perda:	**2,9**	**0**	**12,9**

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

P1: Que ocorre se utilizar a média ao invés da simples soma?

Exemplo – Função de Perda de Articulação

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$ são

gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**
Perda:	**2,9**	**0**	**12,9**

A perda de articulação tem a seguinte forma:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

P1: Que ocorre se utilizar a média ao invés da simples soma?

$$L = \frac{1}{N} \sum_{i=1}^{N} L_i$$

L = (2,9 + 0 + 12,9) / 3 = **5,3**

Função *Softmax*

A ***função softmax*** é uma função que recebe como entrada um vetor e transforma essa entrada em uma ***distribuição de probabilidade***

Função *Softmax*

A ***função softmax*** é uma função que recebe como entrada um vetor e transforma essa entrada em uma ***distribuição de probabilidade***

Seja $f_j(x)$ uma estimativa da probabilidade que x pertença a classe j

Função *Softmax*

A ***função softmax*** é uma função que recebe como entrada um vetor e transforma essa entrada em uma ***distribuição de probabilidade***

Seja $f_j(x)$ uma estimativa da probabilidade que x pertença a classe j

A ***função softmax*** sobre o vetor de "scores" $(s_1, \dots, s_k)$ é dada por

$$f_j(x) = \frac{\exp(s_j)}{\exp(s_1) + \exp(s_2) + \cdots + \exp(s_k)} \quad \text{sendo que} \quad \sum_{j=1}^{k} f_j(x) = 1$$

Função *Softmax*

A ***função softmax*** é uma função que recebe como entrada um vetor e transforma essa entrada em uma ***distribuição de probabilidade***

Seja $f_j(x)$ uma estimativa da probabilidade que x pertença a classe j

A ***função softmax*** sobre o vetor de "scores" $(s_1, \ldots, s_k)$ é dada por

$$f_j(x) = \frac{\exp(s_j)}{\exp(s_1) + \exp(s_2) + \cdots + \exp(s_k)} \quad \text{sendo que} \quad \sum_{j=1}^{k} f_j(x) = 1$$

O nome "softmax" indica que se algum "score" s_j for razoavelmente maior que os demais, a saída será próxima de $(0, \ldots, 1, \ldots, 0)$ em que o 1 se encontra na j posição

Classificador *Softmax*

gato	**3,2**
carro	5,1
rã	-1,7

"Scores" ≡ probabilidades logarítmicas não normalizadas das classes

$$\boxed{s = f(x, W)}$$

Classificador *Softmax*

gato	**3,2**
carro	5,1
rã	-1,7

"Scores" $\equiv$ probabilidades logarítmicas não normalizadas das classes

$$\boxed{P(Y = k|X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}}} \text{ em que } \boxed{s = f(x, W)}$$

Classificador *Softmax*

gato	**3,2**
carro	5,1
rã	-1,7

"Scores" $\equiv$ probabilidades logarítmicas não normalizadas das classes

$$\boxed{P(Y = k|X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}}} \quad \text{em que} \quad \boxed{s = f(x, W)}$$

Deseja-se maximizar a log-verossimilhança, ou ainda, **minimizar a função de log-verossimilhança negativa** da classe correta (considerando uma função de perda)

$$\boxed{L_i = -\log P(Y = y_i \mid X = x_i)}$$

Classificador *Softmax*

gato	**3,2**
carro	5,1
rã	-1,7

"Scores" ≡ probabilidades logarítmicas não normalizadas das classes

$$\boxed{P(Y = k|X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}}} \text{ em que } \boxed{s = f(x, W)}$$

Deseja-se maximizar a log-verossimilhança, ou ainda, **minimizar a função de log-verossimilhança negativa** da classe correta (considerando uma função de perda)

$$\boxed{L_i = -\log P(Y = y_i \mid X = x_i)}$$

Portanto

$$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

Exemplo – Classificador *Softmax*

gato	**3,2**
carro	5,1
rã	-1,7

probabilidades log
não normalizadas

Exemplo – Classificador *Softmax*

	probabilidades log não normalizadas	exp	probabilidades não normalizadas
gato	**3,2**		**24,5**
carro	5,1	→	164,0
rã	-1,7		0,18

Exemplo – Classificador *Softmax*

Função *Softmax*

	probabilidades log não normalizadas		probabilidades não normalizadas		probabilidades
gato	**3,2**		**24,5**		**0,130**
carro	5,1	**exp** →	164,0	**normaliza** →	0,869
rã	-1,7		0,18		0,001

Exemplo – Classificador *Softmax*

$$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

Função *Softmax*

	probabilidades log não normalizadas	exp →	probabilidades não normalizadas	normaliza →	probabilidades
gato	**3,2**		**24,5**		**0,130**
carro	5,1		164,0		0,869
rã	-1,7		0,18		0,001

$L_i = -\log(0{,}13)$
= 0,89

Exemplo de Código – Função de Perda de Articulação

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Código em Python (usando numpy)

```
def L_i_vectorized(x, y, W):
  scores = W.dot(x)
  margins = np.maximum(0, scores - scores[y] + 1)
  margins[y] = 0
  loss_i = np.sum(margins)
  return loss_i
```

Problema com Cálculo da Perda

$$f(x, W) = Wx$$

$$L = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq y_i} \max\left(0, f(x_i, W)_j - f(x_i, W)_{y_i} + 1\right)$$

Problema com Cálculo da Perda

$$f(x, W) = Wx$$

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq y_i} \max\left(0, f(x_i, W)_j - f(x_i, W)_{y_i} + 1\right)$$

Problema com Cálculo da Perda

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$

são

gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**
Perda:	**2,9**	**0**	**12,9**

$$\boxed{L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)}$$

Antes:

= max(0, 1,3 – 4,9 + 1) +
max(0, 2,0 – 4,9 + 1)
= max(0, -2,6) + max(0, -1,9)
= 0 + 0
= 0

Problema com Cálculo da Perda

Suponha: 3 imagens treino e 3 classes

Para algum W, “scores” $s = f(x, W) = Wx$

são

gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**
Perda:	**2,9**	**0**	**12,9**

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Antes:

= max(0, 1,3 – 4,9 + 1) +
max(0, 2,0 – 4,9 + 1)
= max(0, -2,6) + max(0, -1,9)
= 0 + 0
= 0

Com W duas vezes maior:

= max(0, 2,6 – 9,8 + 1) +
max(0, 4,0 – 9,8 + 1)
= max(0, -6,2) + max(0, -4,8)
= 0 + 0
= 0

Problema com Cálculo da Perda

Suponha: 3 imagens treino e 3 classes

Para algum W, "scores" $s = f(x, W) = Wx$

são

gato	**3,2**	1,3	2,2
carro	5,1	**4,9**	2,5
rã	-1,7	2,0	**-3,1**
Perda:	**2,9**	**0**	**12,9**

$$\boxed{L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)}$$

Antes:

= max(0, 1,3 – 4,9 + 1) +
max(0, 2,0 – 4,9 + 1)
= max(0, -2,6) + max(0, -1,9)
= 0 + 0
= 0

Com W duas vezes maior:

= max(0, 2,6 – 9,8 + 1) +
max(0, 4,0 – 9,8 + 1)
= max(0, -6,2) + max(0, -4,8)
= 0 + 0
= 0

=

Problema com Cálculo da Perda

$$f(x, W) = Wx$$

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq y_i} \max\big(0, f(x_i, W)_j - f(x_i, W)_{y_i} + 1\big)$$

Essa perda tem uma dependência linear de $||w^j||_2$ e essa dependência é geralmente negativa, portanto, a minimização do risco tende a fazer crescer $||w^j||_2$

Problema com Cálculo da Perda

$$f(x, W) = Wx$$

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq y_i}\max\big(0, f(x_i, W)_j - f(x_i, W)_{y_i} + 1\big)$$

Essa perda tem uma dependência linear de $||w^j||_2$ e essa dependência é geralmente negativa, portanto, a minimização do risco tende a fazer crescer $||w^j||_2$

Pode-se tentar corrigir isso usando um termo de regularização $\lambda||W||_2$ em que a norma da matrix $||\cdot||_2$ representa soma dos quadrados de seus elementos

Problema com Cálculo da Perda

$$f(x, W) = Wx$$

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq y_i} \max\left(0, f(x_i, W)_j - f(x_i, W)_{y_i} + 1\right)$$

Essa perda tem uma dependência linear de $||w^j||_2$ e essa dependência é geralmente negativa, portanto, a minimização do risco tende a fazer crescer $||w^j||_2$

Pode-se tentar corrigir isso usando um termo de regularização $\lambda||W||_2$ em que a norma da matrix $||\cdot||_2$ representa soma dos quadrados de seus elementos

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq y_i} \max\left(0, f(x_i, W)_j - f(x_i, W)_{y_i} + 1\right) \;\; + \boldsymbol{\lambda||W||_2}$$

Força ou Peso de Regularização

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq y_i} \max\left(0, f(x_i, W)_j - f(x_i, W)_{y_i} + 1\right) \; + \; \lambda R(W)$$

Força ou Peso de Regularização

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq y_i}\max\left(0, f(x_i,W)_j - f(x_i,W)_{y_i} + 1\right) \;+\; \boxed{\lambda R(W)}$$

λ= força de regularização (hiperparâmetro)

Força ou Peso de Regularização

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq y_i} \max\left(0, f(x_i,W)_j - f(x_i,W)_{y_i} + 1\right) \;+\; \boxed{\lambda R(W)}$$

λ= força de regularização (hiperparâmetro)

Regularizações comuns:

- L2 (Ridge) $R(W) = \sum_k \sum_l W_{k,l}^2$

- L1 (Lasso) $R(W) = \sum_k \sum_l \left|W_{k,l}\right|$

- L1 + L2 (*Elastic net*) $R(W) = \sum_k \sum_l \beta W_{k,l}^2 + \left|W_{k,l}\right|$

- Técnicas mais recentes: ***Dropout***, ...

Recapitulando

- Tem-se um conjunto de dados (x, y)

Recapitulando

- Tem-se um conjunto de dados (*x*, *y*)
- Uma função de **“score”** p.ex. $s = f(x; W) = Wx$

Recapitulando

- Tem-se um conjunto de dados (*x*, *y*)
- Uma função de **“score”** p.ex. $s = f(x; W) = Wx$
- Uma função de **perda**

Log-Ver Neg $$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

Hinge $$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Recapitulando

- Tem-se um conjunto de dados (*x*, *y*)
- Uma função de **“score”** p.ex. $s = f(x; W) = Wx$
- Uma função de **perda**

Log-Ver Neg $$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

Hinge $$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Perda Total $$L = \frac{1}{N} \sum_{i=1}^{N} L_i + R(W)$$

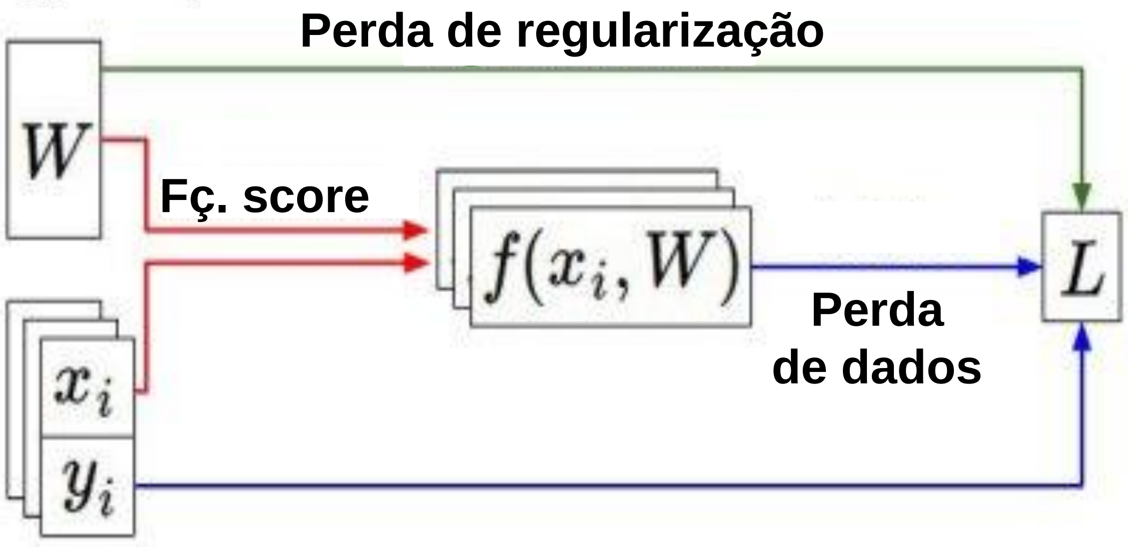
Recapitulando

- Tem-se um conjunto de dados (*x*, *y*)
- Uma função de **“score”** p.ex. $s = f(x; W) = Wx$
- Uma função de **perda**

Log-Ver Neg $$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

Hinge $$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

Perda Total $$L = \frac{1}{N} \sum_{i=1}^{N} L_i + R(W)$$



Grafo de Computação da Função de Perda