UNIVERSIDADE FEDERAL DO CEARÁ

DEPARTAMENTO DE COMPUTAÇÃO

Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2024

Agenda

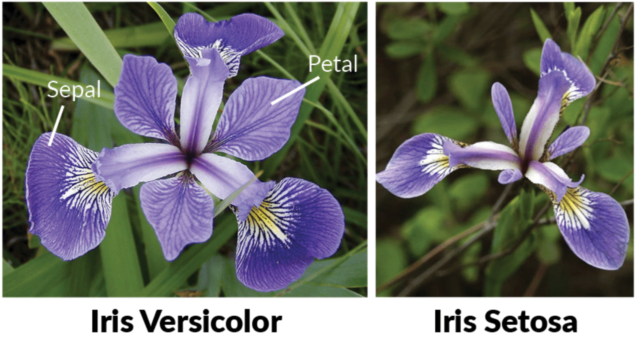
Classificação

Tarefa de classificação

Relaciona vetores de entrada a um número finito de rótulos/categorias/classes de saída.

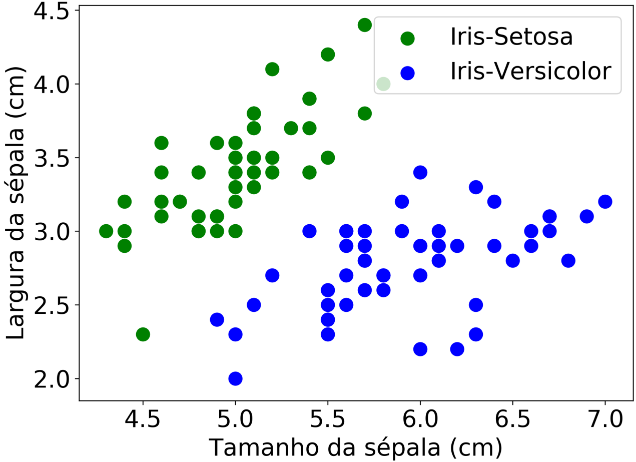
- **Classificação binária**: Somente duas classes (sim/não, positivo/negativo, gato/cachorro, etc.)
- **Classificação multiclasse**: Mais de duas classes (dígitos, letras, raças de cachorro, marcas de carro, etc.)

Classificação binária



- **Problema**: Como classificar automaticamente flores da espécie íris entre Setosa e Versicolor a partir de medidas de suas sépalas?

Classificação binária



- **Ideia**: Podemos utilizar um modelo de regressão linear nessa tarefa de classificação?

Classificação binária

- Convertemos as saídas categóricas (“setosa” ou “versicolor”) em números: -1 ou 1.

Classificação binária

- Convertemos as saídas categóricas ("setosa" ou "versicolor") em números: -1 ou 1.
- **Problema**: O modelo $\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$ retorna valores reais.

Classificação binária

- Convertemos as saídas categóricas ("setosa" ou "versicolor") em números: -1 ou 1.
- **Problema**: O modelo $\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$ retorna valores reais.
- **Ideia**: Modificar a saída para $\hat{y}_i = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i)$, em que:

$$\text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i) = \begin{cases} -1 & \text{, se } \boldsymbol{w}^\top \boldsymbol{x}_i < 0 \\ 1 & \text{, se } \boldsymbol{w}^\top \boldsymbol{x}_i \geq 0 \end{cases}.$$

Classificação binária

- Convertemos as saídas categóricas ("setosa" ou "versicolor") em números: -1 ou 1.
- **Problema**: O modelo $\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$ retorna valores reais.
- **Ideia**: Modificar a saída para $\hat{y}_i = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i)$, em que:

$$\text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i) = \begin{cases} -1 & \text{, se } \boldsymbol{w}^\top \boldsymbol{x}_i < 0 \\ 1 & \text{, se } \boldsymbol{w}^\top \boldsymbol{x}_i \geq 0 \end{cases}.$$

- **Problema**: Como modificar a regra de atualização dos parâmetros, dado que a função $\text{sign}(\cdot)$ não é diferenciável?

Classificação binária

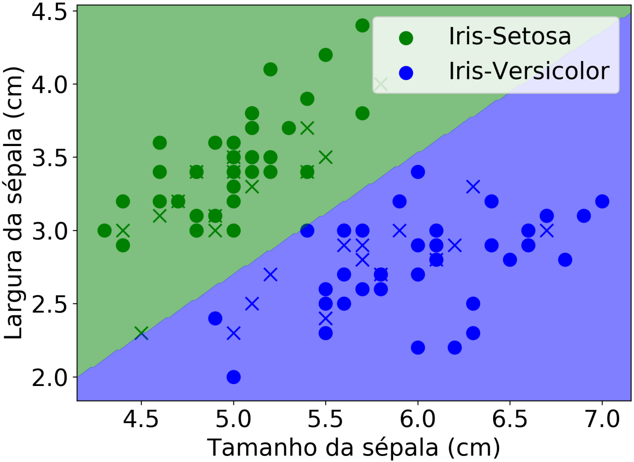
- Convertemos as saídas categóricas (“setosa” ou “versicolor”) em números: -1 ou 1.
- **Problema**: O modelo $\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$ retorna valores reais.
- **Ideia**: Modificar a saída para $\hat{y}_i = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i)$, em que:

$$\text{sign}(\boldsymbol{w}^\top \boldsymbol{x}_i) = \begin{cases} -1 & \text{, se } \boldsymbol{w}^\top \boldsymbol{x}_i < 0 \\ 1 & \text{, se } \boldsymbol{w}^\top \boldsymbol{x}_i \geq 0 \end{cases}.$$

- **Problema**: Como modificar a regra de atualização dos parâmetros, dado que a função $\text{sign}(\cdot)$ não é diferenciável?
- **Ideia**: Vamos usar a função $\text{sign}(\cdot)$ somente na predição do modelo.

Classificação binária

- Solução via OLS: $\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$
- Classificação binária (70% para treinamento e 30% para teste):



Agenda

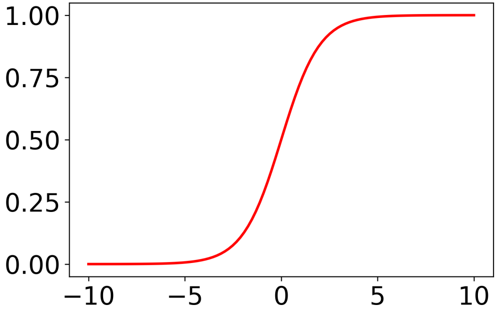
Classificação binária

- **Ideia**: Trocar a função $\text{sign}(\cdot)$ por uma função diferenciável entre 0 e 1.

Classificação binária

- **Ideia**: Trocar a função $\text{sign}(\cdot)$ por uma função diferenciável entre 0 e 1.
- **Função logística (sigmóide)**:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$



Classificação binária

Regressão logística

- Apesar do nome, é um método de **classificação**.
- Usa uma **função logística** na saída do modelo linear:

$$\hat{y}_i = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i), \quad \sigma(z) = \frac{1}{1 + \exp(-z)}.$$

- A função logística é definida no intervalo $[0, 1]$, possuindo interpretação probabilística.
- $\sigma(z)$ é facilmente **diferenciável**:

$$\frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} = \sigma(z)(1 - \sigma(z)).$$

Regressão logística binária

- **Problema**: Como modelar probabilisticamente os dados a partir da função logística?

Regressão logística binária

- **Problema**: Como modelar probabilisticamente os dados a partir da função logística?

Distribuição de Bernoulli

- Seja uma moeda potencialmente injusta (cara (1) e coroa (0)):

$$P(y = 1|q) = q,$$
$$P(y = 0|q) = 1 - q.$$

- A Distribuição de Bernoulli é então definida por:

$$p(y|q) = q^y (1 - q)^{1-y}.$$

Regressão logística binária

- **Problema**: Como modelar probabilisticamente os dados a partir da função logística?

Verossimilhança de Bernoulli

- Considerando duas classes, 0 e 1, temos:

$$
\begin{aligned}
P(y=1|\boldsymbol{x},\boldsymbol{w}) &= \sigma(\boldsymbol{w}^\top\boldsymbol{x}),\\
P(y=0|\boldsymbol{x},\boldsymbol{w}) &= 1-\sigma(\boldsymbol{w}^\top\boldsymbol{x}).
\end{aligned}
$$

- A verossimilhança de Bernoulli é então definida por:

$$
p(y|\boldsymbol{x},\boldsymbol{w}) = \sigma(\boldsymbol{w}^\top\boldsymbol{x})^y(1-\sigma(\boldsymbol{w}^\top\boldsymbol{x}))^{1-y}.
$$

Regressão logística binária

- **Problema**: Qual será a nova função custo?

Regressão logística binária

- **Problema**: Qual será a nova função custo?
- **Ideia**: Escolher o negativo da **log-verossimilhança**:

$$\mathcal{J}(\boldsymbol{w}) = -\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$$

$$\mathcal{J}(\boldsymbol{w}) = -\log \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{w})$$

$$\mathcal{J}(\boldsymbol{w}) = -\log \prod_{i=1}^{N} \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)^{y_i}(1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))^{1-y_i}$$

$$\mathcal{J}(\boldsymbol{w}) = -\sum_{i=1}^{N} \left[y_i \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1 - y_i) \log(1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)) \right].$$

Regressão logística binária

Cross entropy loss

- Definida por:

$$\mathcal{J}(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1 - y_i)\log(1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))\right].$$

- Precisamos calcular o gradiente da função custo para atualizar os parâmetros do modelo:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}}.$$

Regressão logística binária

- Derivando em relação a $\boldsymbol{w}$, temos:

$$
\begin{aligned}
\mathcal{J}(\boldsymbol{w}) &= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))\right], \\
\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \frac{1}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)}\frac{\partial \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial \boldsymbol{w}} - (1-y_i)\frac{1}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)}\frac{\partial \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)}{\partial \boldsymbol{w}}\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \frac{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)(1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)}\boldsymbol{x}_i - (1-y_i)\frac{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)(1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)}\boldsymbol{x}_i\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i(1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))\boldsymbol{x}_i - (1-y_i)\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\boldsymbol{x}_i - y_i\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i + y_i\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\boldsymbol{x}_i\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}(y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i))\boldsymbol{x}_i \\
&= -\frac{1}{N}\sum_{i=1}^{N} e_i \boldsymbol{x}_i.
\end{aligned}
$$

Regressão logística binária

- Com o gradiente $\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}}$, atualizamos o modelo via GD ou SGD.

Gradiente Descendente (GD)

- Regra de atualização:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i(t-1)\boldsymbol{x}_i$$

Gradiente Descendente Estocástico (SGD)

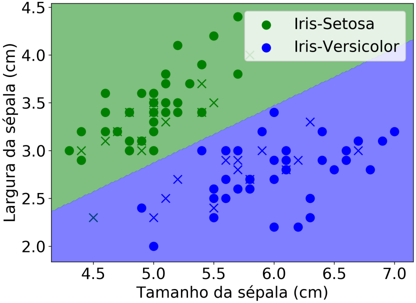
- Regra de atualização:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha\, e_i(t-1)\boldsymbol{x}_i$$

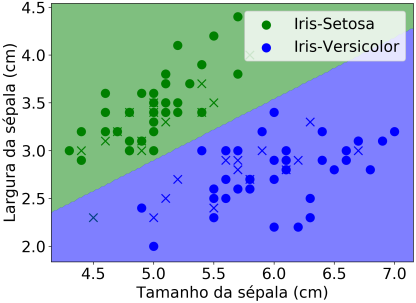
- Lembrando que na regressão logística temos:

$$e_i(t) = y_i - \sigma(\boldsymbol{w}(t)^\top \boldsymbol{x}_i)$$

Exemplo de classificação (dados separáveis linearmente)

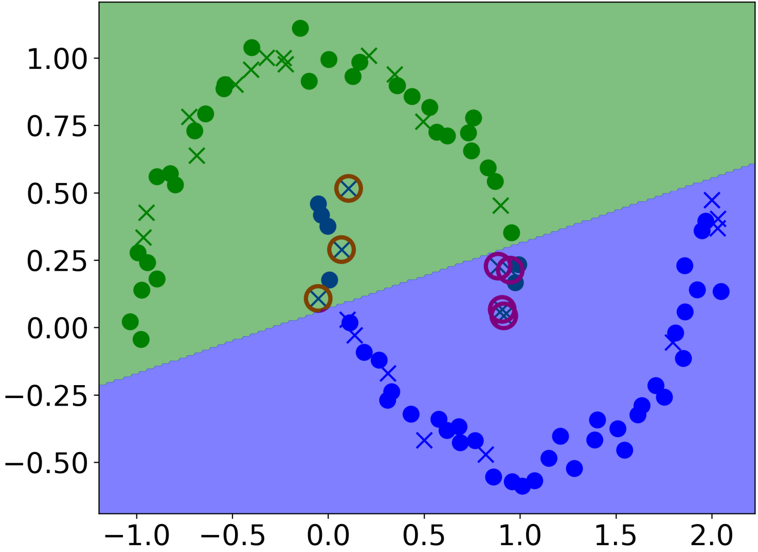


Regressão logística via GD



Regressão logística via SGD

Exemplo de classificação (dados não separáveis linearmente)



Regressão logística binária

Extensões da regressão logística

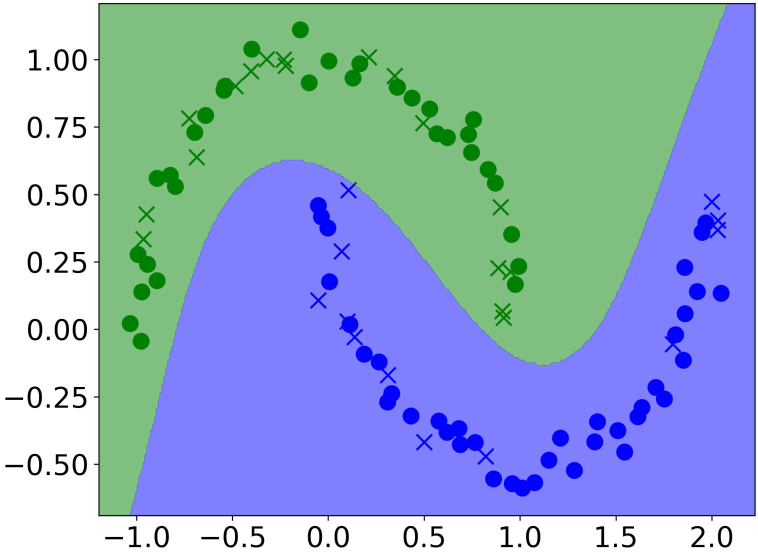
- Novos **atributos não-lineares** $(x_i^2, x_i^3, \cdots)$ podem ser incluídos para obter um **classificador não-linear**.

Regressão logística binária

Extensões da regressão logística

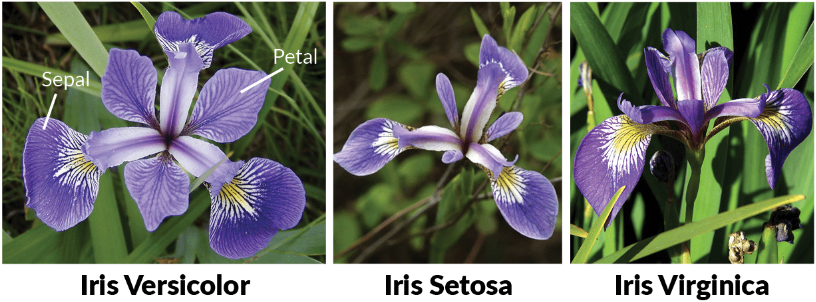
- Novos **atributos não-lineares** $(x_i^2, x_i^3, \cdots)$ podem ser incluídos para obter um **classificador não-linear**.
- Modelos de regressão logística também podem ser **regularizados**.
  - Inclui na função custo o termo: $+\lambda\|\boldsymbol{w}\|^2$.
  - Inclui na regra de atualização o termo: $-\lambda\boldsymbol{w}(t-1)$.

Exemplo de classificação com atributos polinomiais



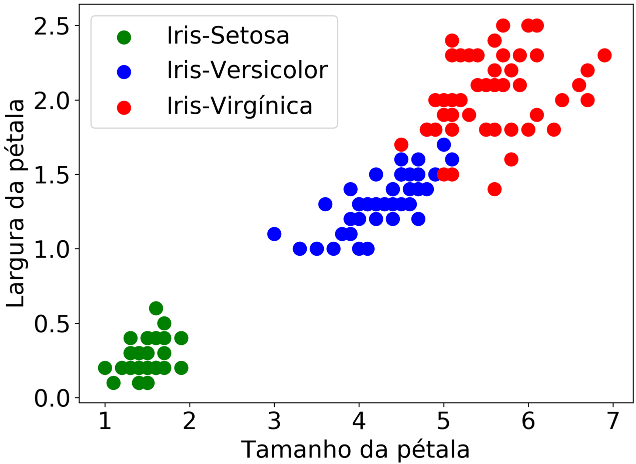
Agenda

Classificação multiclasse



- **Problema**: Como classificar automaticamente flores da espécie íris entre Setosa, Versicolor e Virgínica a partir de medidas de suas pétalas?

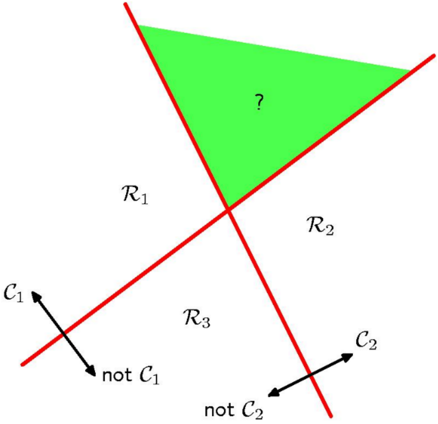
Classificação multiclasse



- **Problema**: Como representamos as classes na saída do modelo?

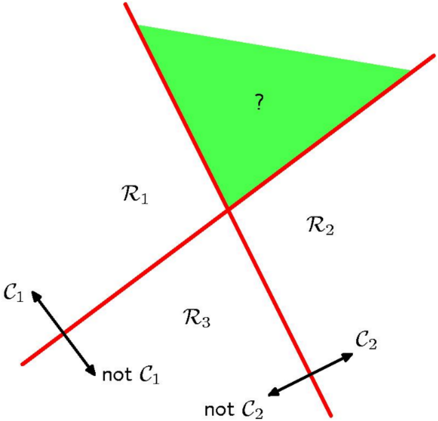
Classificação multiclasse

- **Ideia**: $K - 1$ classificações binárias **one vs all**:



Classificação multiclasse

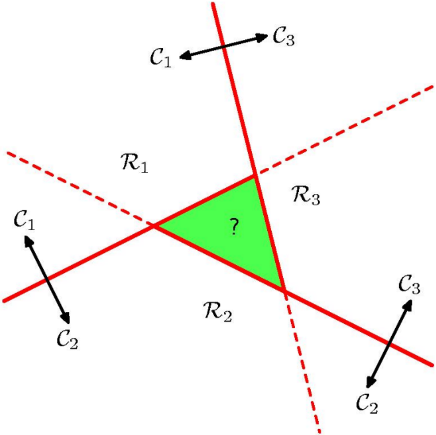
- **Ideia**: $K-1$ classificações binárias **one vs all**:



- **Problema**: Regiões não associadas a uma única classe.

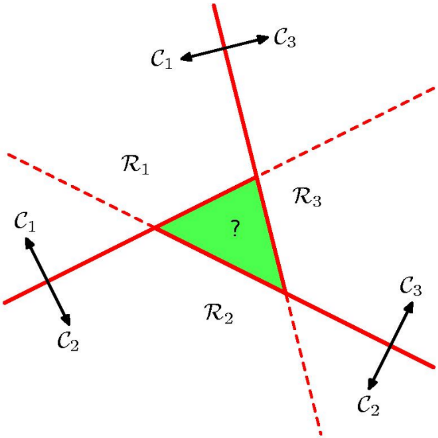
Classificação multiclasse

- **Ideia**: $K(K-1)/2$ classificações binárias **one vs one**:



Classificação multiclasse

- **Ideia**: $K(K-1)/2$ classificações binárias **one vs one**:



- **Problema**: Regiões não associadas a uma única classe.

Classificação multiclasse

One hot encoding (1-of-K encoding)

- A saída do modelo é um vetor de K elementos ($K =$ número de classes).
- O vetor de saída desejado $\boldsymbol{y}_i$ consiste em um vetor de $K-1$ zeros e um valor 1 na k-ésima posição associada à k-ésima classe.
- **Exemplo**: $\boldsymbol{y}_i = [1\ 0\ 0]^\top$, ou $\boldsymbol{y}_i = [0\ 1\ 0]^\top$, ou $\boldsymbol{y}_i = [0\ 0\ 1]^\top$.

Classificação multiclasse

One hot encoding (1-of-K encoding)

- A saída do modelo é um vetor de K elementos (K = número de classes).
- O vetor de saída desejado $\boldsymbol{y}_i$ consiste em um vetor de $K-1$ zeros e um valor 1 na k-ésima posição associada à k-ésima classe.
- **Exemplo**: $\boldsymbol{y}_i = [1\ 0\ 0]^\top$, ou $\boldsymbol{y}_i = [0\ 1\ 0]^\top$, ou $\boldsymbol{y}_i = [0\ 0\ 1]^\top$.

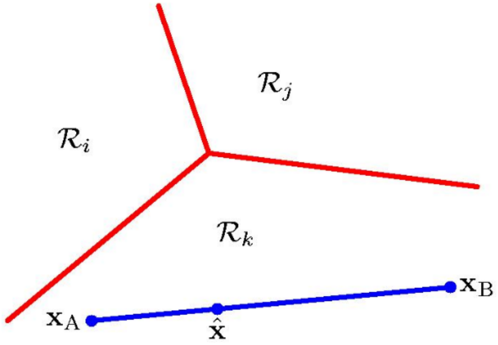
Discriminante linear

- Dado um total de K classes, a classe k_* predita para o padrão $\boldsymbol{x}_*$ é dada por:

$$k_* = \arg \max_{1 \leq k \leq K} \hat{y}_k.$$

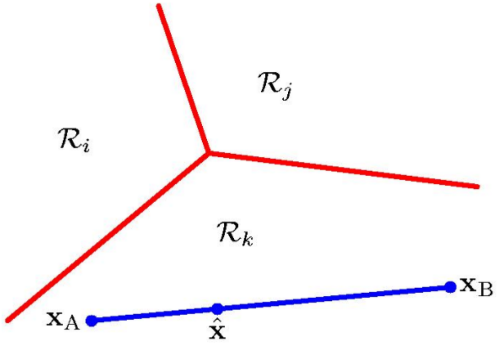
Classificação multiclasse

- As regiões definidas por um discriminante linear são **convexas**:



Classificação multiclasse

- As regiões definidas por um discriminante linear são **convexas**:



- **Problema**: Notação do modelo com múltiplas saídas?

Regressão multivariada

- Nova notação matricial:

$$\hat{\boldsymbol{y}}_i = \boldsymbol{W}^\top \boldsymbol{x}_i,$$
$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W},$$

→ $\boldsymbol{W} \in \mathbb{R}^{D \times K}$ é a matriz de parâmetros do modelo.
→ $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ é a coleção de entradas do modelo.
→ $\hat{\boldsymbol{Y}} \in \mathbb{R}^{N \times K}$ é a coleção de saídas do modelo.

Regressão multivariada

OLS para regressão multivariada (múltiplas saídas)

- **Função custo**:

$$\mathcal{J}(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_F^2 = \frac{1}{2}\sum_{i=1}^{N}\|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\|^2 = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}|y_{ik} - \hat{y}_{ik}|^2,$$

  em que $\|\cdot\|_F$ é a **Norma de Frobenius**.

- **Solução analítica**:

$$\boldsymbol{W} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}.$$

Regressão multivariada

Gradiente Descendente para múltiplas saídas

- Regra de atualização:

$$\boldsymbol{w}_k(t) = \boldsymbol{w}_k(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_{ik}(t-1)\boldsymbol{x}_i$$

Gradiente Descendente Estocástico para múltiplas saídas

- Regra de atualização:

$$\boldsymbol{w}_k(t) = \boldsymbol{w}_k(t-1) + \alpha\, e_{ik}(t-1)\boldsymbol{x}_i$$

- Note que:

$\rightarrow$ $e_{ik} = y_{ik} - \hat{y}_{ik}$

$\rightarrow$ $\boldsymbol{W} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_k \cdots \boldsymbol{w}_K],\ \boldsymbol{w}_k \in \mathbb{R}^D$

Classificação multiclasse

Regressão logística multiclasse

- A coluna $\boldsymbol{w}_k$ da matriz $\boldsymbol{W}$ está associada à classe k.

Classificação multiclasse

Regressão logística multiclasse

- A coluna $\boldsymbol{w}_k$ da matriz $\boldsymbol{W}$ está associada à classe k.
- Para a saída do modelo, usamos a função **softmax**:

$$\hat{y}_{ik} = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)}{\sum_{j=1}^{K} \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i)}, \quad 1 \leq k \leq K.$$

Classificação multiclasse

Regressão logística multiclasse

- A coluna $\boldsymbol{w}_k$ da matriz $\boldsymbol{W}$ está associada à classe k.
- Para a saída do modelo, usamos a função **softmax**:

$$\hat{y}_{ik} = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)}{\sum_{j=1}^{K} \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i)}, \quad 1 \leq k \leq K.$$

- Interpretação probabilística: $\hat{y}_{ik} = p(y_{ik}|\boldsymbol{x}_i, \boldsymbol{W}) \in [0, 1]$.
- Também chamada de **regressão softmax** ou **regressão logística multinomial**.

Classificação multiclasse

Regressão logística multiclasse

- A coluna $\boldsymbol{w}_k$ da matriz $\boldsymbol{W}$ está associada à classe k.
- Para a saída do modelo, usamos a função **softmax**:

$$\hat{y}_{ik} = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)}{\sum_{j=1}^{K} \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i)}, \quad 1 \leq k \leq K.$$

- Interpretação probabilística: $\hat{y}_{ik} = p(y_{ik} | \boldsymbol{x}_i, \boldsymbol{W}) \in [0, 1]$.
- Também chamada de **regressão softmax** ou **regressão logística multinomial**.

- **Problema**: Qual será a nova função custo?

Regressão logística multiclasse

Multiclass cross-entropy

- Função custo para regressão logística multiclasse:

$$\mathcal{J}(\boldsymbol{W}) = -\frac{1}{N}\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W})$$

$$\mathcal{J}(\boldsymbol{W}) = -\frac{1}{N}\log \prod_{i=1}^{N}\prod_{k=1}^{K} p(y_{ik}|\boldsymbol{x}_i, \boldsymbol{W})^{y_{ik}}$$

$$\mathcal{J}(\boldsymbol{W}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log \hat{y}_{ik}.$$

- Precisamos calcular o gradiente da função custo para atualizar os parâmetros do modelo:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \alpha\frac{\partial \mathcal{J}(\boldsymbol{W})}{\partial \boldsymbol{W}}, \text{ ou } \boldsymbol{w}_k \leftarrow \boldsymbol{w}_k - \alpha\frac{\partial \mathcal{J}(\boldsymbol{W})}{\partial \boldsymbol{w}_k}, \forall k.$$

Regressão logística multiclasse

- As derivadas em relação aos parâmetros são dadas por:

$$\mathcal{J}(\boldsymbol{W}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log \hat{y}_{ij},$$

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}_k} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} \frac{y_{ij}}{\hat{y}_{ij}}\frac{\partial \hat{y}_{ij}}{\partial \boldsymbol{w}_k}, \text{ em que:}$$

$$\frac{\partial \hat{y}_{ik}}{\partial \boldsymbol{w}_k} = \frac{\partial}{\partial \boldsymbol{w}_k}\left[\frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}_i)}{\sum_{c=1}^{K}\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}\right] = (\hat{y}_{ik} - \hat{y}_{ik}^2)\boldsymbol{x}_i, \quad \text{se } j = k,$$

$$\frac{\partial \hat{y}_{ij}}{\partial \boldsymbol{w}_k} = \frac{\partial}{\partial \boldsymbol{w}_k}\left[\frac{\exp(\boldsymbol{w}_j^\top \boldsymbol{x}_i)}{\sum_{c=1}^{K}\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}\right] = -\hat{y}_{ij}\hat{y}_{ik}\boldsymbol{x}_i, \quad \text{se } j \neq k,$$

ou seja: $\frac{\partial \hat{y}_{ij}}{\partial \boldsymbol{w}_k} = [\delta(j,k)\hat{y}_{ik} - \hat{y}_{ij}\hat{y}_{ik}]\boldsymbol{x}_i, \quad \delta(j,k) = \begin{cases} 1, & j = k, \\ 0, & j \neq k \end{cases}.$

Regressão logística multiclasse

- Substituindo na derivada original:

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}_k} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\frac{y_{ij}}{\hat{y}_{ij}}[\delta(j,k)\hat{y}_{ij} - \hat{y}_{ij}\hat{y}_{ik}]\boldsymbol{x}_i$$

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}_k} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}[\delta(j,k) - \hat{y}_{ik}]\boldsymbol{x}_i$$

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}_k} = -\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{K} y_{ij}\delta(j,k) - \hat{y}_{ik}\underbrace{\sum_{j=1}^{K} y_{ij}}_{=1}\right]\boldsymbol{x}_i$$

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}_k} = -\frac{1}{N}\sum_{i=1}^{N}[y_{ik} - \hat{y}_{ik}]\,\boldsymbol{x}_i = -\frac{1}{N}\sum_{i=1}^{N} e_{ik}\boldsymbol{x}_i.$$

- Note que a soma dos elementos do vetor $\boldsymbol{y}_i$ é igual a 1.

Regressão logística multiclasse

- Com os gradientes $\frac{\partial \mathcal{J}(\boldsymbol{W})}{\partial \boldsymbol{w}_k}$, atualizamos o modelo via GD/SGD.

Gradiente Descendente para múltiplas saídas

- Regra de atualização:

$$\boldsymbol{w}_k(t) = \boldsymbol{w}_k(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_{ik}(t-1)\boldsymbol{x}_i$$

Gradiente Descendente Estocástico para múltiplas saídas

- Regra de atualização:

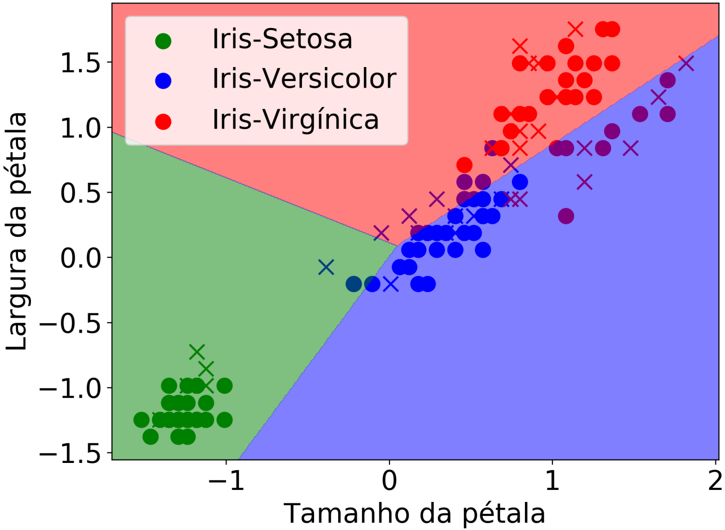
$$\boldsymbol{w}_k(t) = \boldsymbol{w}_k(t-1) + \alpha\, e_{ik}(t-1)\boldsymbol{x}_i$$

- Lembrando que na regressão logística multiclasse temos:

$$e_{ik}(t) = y_{ik} - \frac{\exp(\boldsymbol{w}_k(t)^\top \boldsymbol{x}_i)}{\sum_{j=1}^{K} \exp(\boldsymbol{w}_j(t)^\top \boldsymbol{x}_i)}.$$

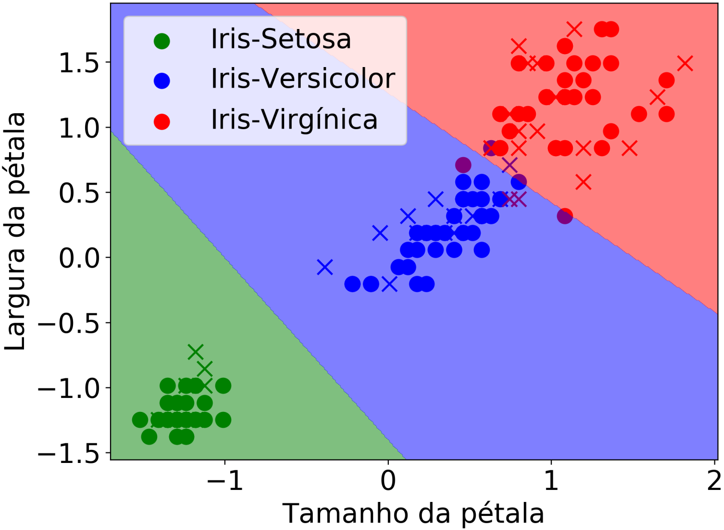
Classificação multiclasse

Regressão linear "ingênua" (OLS) - 72.73% de acurácia no teste



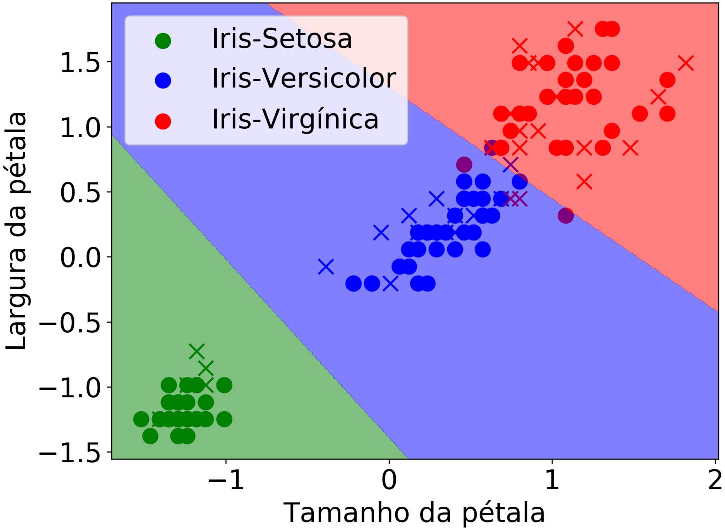
Classificação multiclasse

Regressão logística (GD) - 93.18% de acurácia no teste



Classificação multiclasse

Regressão logística (SGD) - 93.18% de acurácia no teste



Entropia cruzada - visão alternativa

- Buscamos minimizar a discrepância entre a distribuição empírica dos dados $p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x})$ e a distribuição do modelo $p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})$.

Entropia cruzada - visão alternativa

- Buscamos minimizar a discrepância entre a distribuição empírica dos dados $p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x})$ e a distribuição do modelo $p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})$.
- **Divergência de Kullback-Leibler (KL)**: quantifica estatisticamente a diferença entre duas distribuições:

$$
\begin{aligned}
\mathrm{KL}(p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x})||p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})) &= \sum_i p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i) \log \frac{p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i)}{p_{\boldsymbol{w}}(\boldsymbol{y}_i|\boldsymbol{x}_i)} \\
&= \sum_i p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i) \log p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i) - \sum_i p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i) \log p_{\boldsymbol{w}}(\boldsymbol{y}_i|\boldsymbol{x}_i) \\
&= -\mathcal{H}(p_{\mathcal{D}}(\boldsymbol{y})) + \mathcal{H}(p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x}), p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})).
\end{aligned}
$$

Entropia cruzada - visão alternativa

- Buscamos minimizar a discrepância entre a distribuição empírica dos dados $p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x})$ e a distribuição do modelo $p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})$.
- **Divergência de Kullback-Leibler (KL)**: quantifica estatisticamente a diferença entre duas distribuições:

$$\begin{aligned}
\mathrm{KL}(p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x})||p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})) &= \sum_i p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i)\log\frac{p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i)}{p_{\boldsymbol{w}}(\boldsymbol{y}_i|\boldsymbol{x}_i)} \\
&= \sum_i p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i)\log p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i) - \sum_i p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i)\log p_{\boldsymbol{w}}(\boldsymbol{y}_i|\boldsymbol{x}_i) \\
&= -\mathcal{H}(p_{\mathcal{D}}(\boldsymbol{y})) + \mathcal{H}(p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x}), p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})).
\end{aligned}$$

- Como a entropia $\mathcal{H}(p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x}))$ não depende dos parâmetros $\boldsymbol{w}$, minimizar o KL em relação a $\boldsymbol{w}$ equivale a minimizar a entropia cruzada $\mathcal{H}(p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x}), p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x}))$.

Entropia cruzada - visão alternativa

- Portanto, busca-se minimizar a seguinte função custo em relação aos parâmetros $\boldsymbol{w}$:

$$\begin{aligned}\mathcal{H}(p_{\mathcal{D}}(\boldsymbol{y}|\boldsymbol{x}), p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})) &= -\sum_i p_{\mathcal{D}}(\boldsymbol{y}_i|\boldsymbol{x}_i) \log p_{\boldsymbol{w}}(\boldsymbol{y}_i|\boldsymbol{x}_i) \\ &= -\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\mathcal{D}}}[\log p_{\boldsymbol{w}}(\boldsymbol{y}|\boldsymbol{x})].\end{aligned}$$

- Nota-se que a entropia cruzada é o negativo da log-verossimilhança calculada sobre os dados observados.
- Isso é verdadeiro para qualquer cenário de estimação por máxima verossimilhança (MLE), não somente classificação!

Agenda

Tópicos adicionais

- Representação de atributos categóricos via one hot encoding.
  - → **Exemplo**: Atributo "gênero de filme" (ação, drama ou comédia): $\boldsymbol{x}_i = [1\ 0\ 0]^\top$, ou $\boldsymbol{x}_i = [0\ 1\ 0]^\top$, ou $\boldsymbol{x}_i = [0\ 0\ 1]^\top$.
- Métodos de segunda ordem para regressão logística, como o iteratively reweighted least squares (IRLS) ou o BFGS.
- Generalized linear models (GLMs).
- Regressão ordinal.

Agenda

Referências bibliográficas

- **Cap. 8** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Caps. 2, 4 e 10** - MURPHY, Kevin P. **Probabilistic Machine Learning: An Introduction**, 2021.
- **Cap. 4*** - BISHOP, Christopher M. **Pattern recognition and machine learning**, 2006.