Universidade Federal do Ceará
Departamento de Computação

Aprendizagem de Máquina Probabilística

César Lincoln Cavalcante Mattos

2024

Agenda

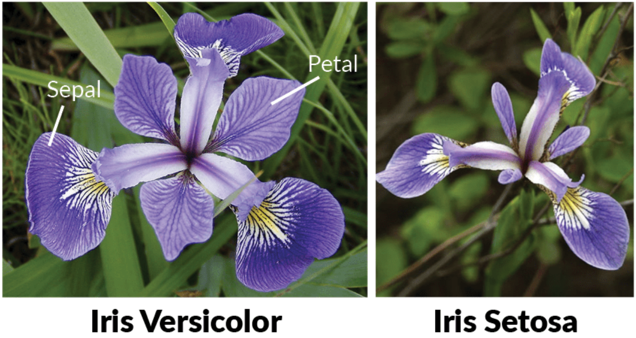
Classificação

Tarefa de classificação

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

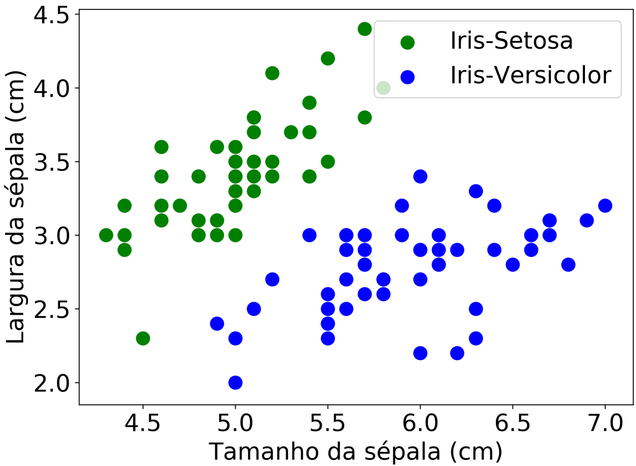
- **Classificação binária**: Somente duas classes (sim/não, positivo/negativo, gato/cachorro, etc.)
- **Classificação multiclasse**: Mais de duas classes (dígitos, letras, raças de cachorro, marcas de carro, etc.)

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 $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i$ retorna valores reais.

Classificação binária

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

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

Classificação binária

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

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

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

Classificação binária

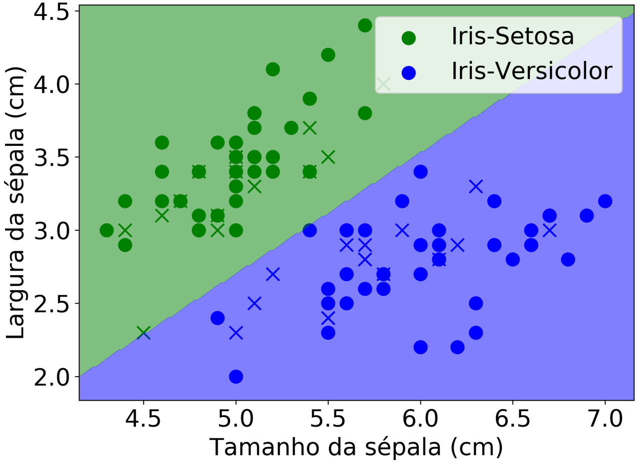
- Convertemos as saídas categóricas ("setosa" ou "versicolor") em números: -1 ou 1.
- **Problema**: O modelo $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i$ retorna valores reais.
- **Ideia**: Modificar a saída para $\hat{y}_i = \text{sign}(\mathbf{w}^\top \mathbf{x}_i)$, em que:

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

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



Agenda

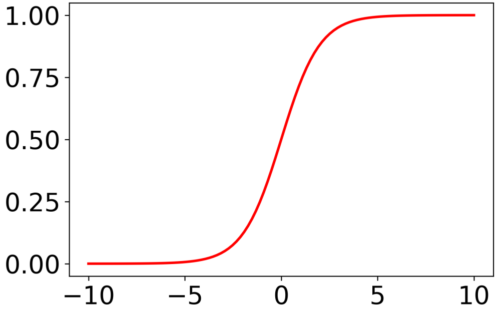
Classificação binária

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

Classificação binária

- **Ideia**: Trocar a função 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:

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

- 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?

Distribuição de Bernoulli

- Considerando duas classes, 0 e 1, temos:

$$P(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}),$$
$$P(y = 0|\mathbf{x}, \mathbf{w}) = 1 - \sigma(\mathbf{w}^\top \mathbf{x}).$$

- A Distribuição de Bernoulli é então definida por:

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

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**:

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

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

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

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

Regressão logística binária

Cross entropy loss

- Definida por:

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

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

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

Regressão logística binária

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

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

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

Regressão logística binária

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

Gradiente Descendente (GD)

- Regra de atualização:

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

Gradiente Descendente Estocástico (SGD)

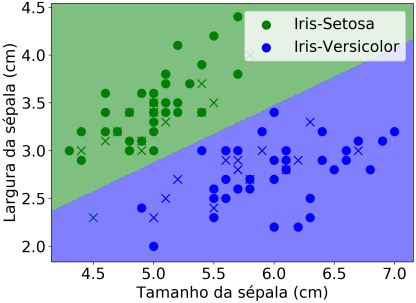
- Regra de atualização:

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

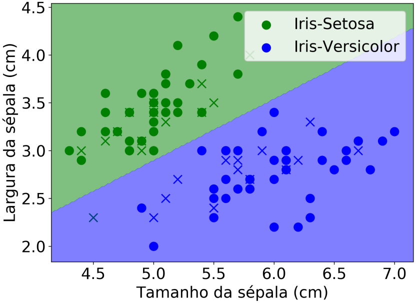
- Lembrando que na regressão logística temos:

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

Exemplo de classificação (dados separáveis linearmente)

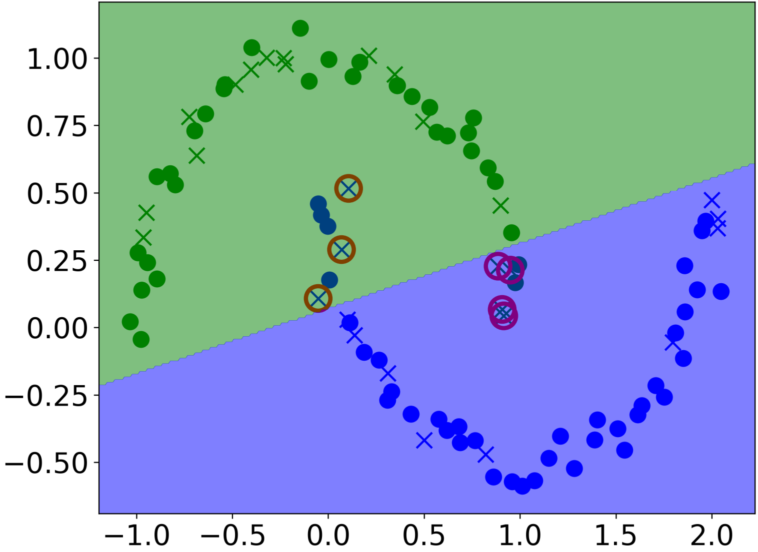


Regressão logística via GD



Regressão logística via SGD

Exemplo de classificação (dados não separáveis linearmente)



Regressão logística binária

Extensões da regressão logística

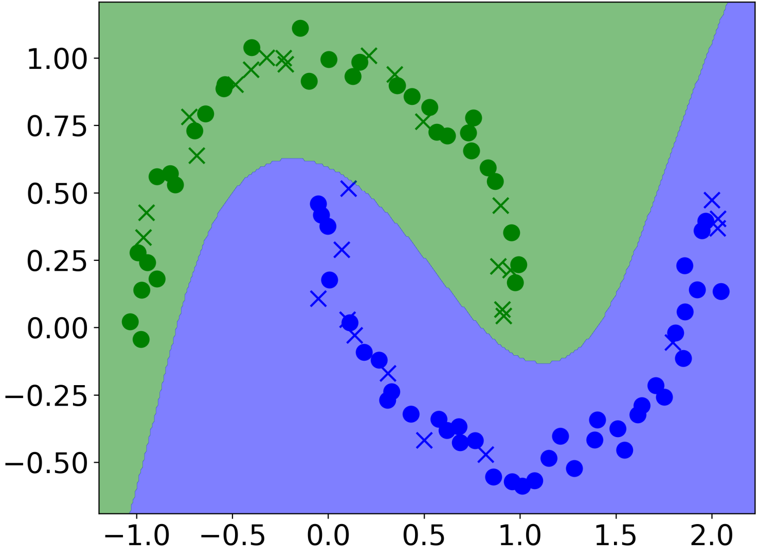
- Novos **atributos não-lineares** $(x_i^2, x_i^3, \cdots)$ podem ser incluídos para obter um **classificador não-linear**.

Regressão logística binária

Extensões da regressão logística

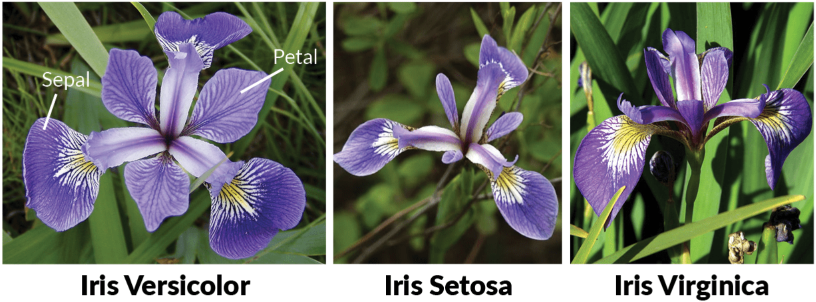
- Novos **atributos não-lineares** $(x_i^2, x_i^3, \cdots)$ podem ser incluídos para obter um **classificador não-linear**.
- Modelos de regressão logística também podem ser **regularizados**.
  - Inclui na função custo o termo: $+\lambda\|\mathbf{w}\|^2$.
  - Inclui na regra de atualização o termo: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

Exemplo de classificação com atributos polinomiais



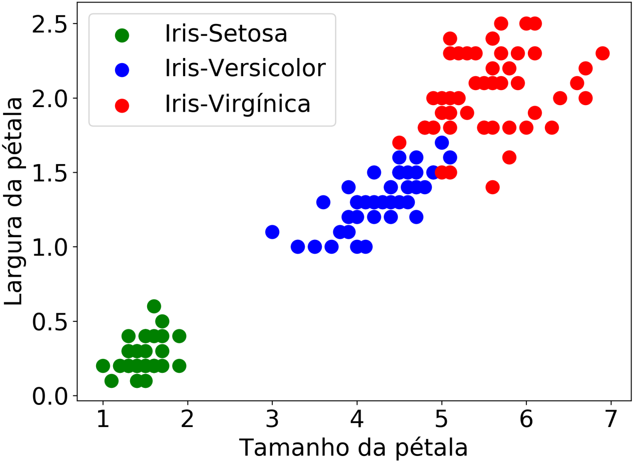
Agenda

Classificação multiclasse



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

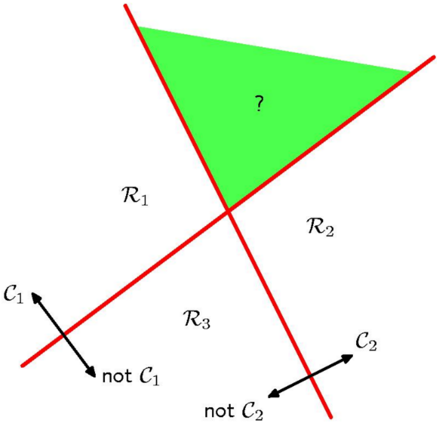
Classificação multiclasse



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

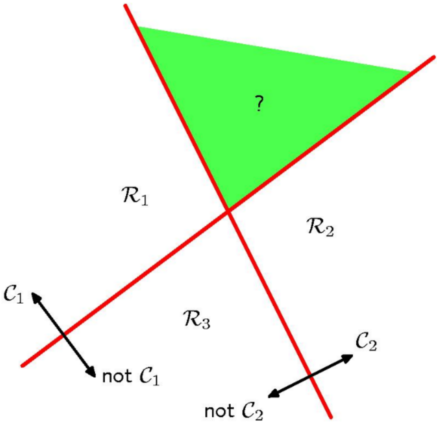
Classificação multiclasse

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



Classificação multiclasse

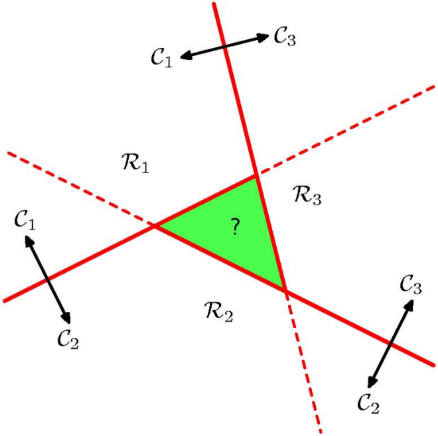
- **Ideia**: $K - 1$ classificações binárias **one vs all**:



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

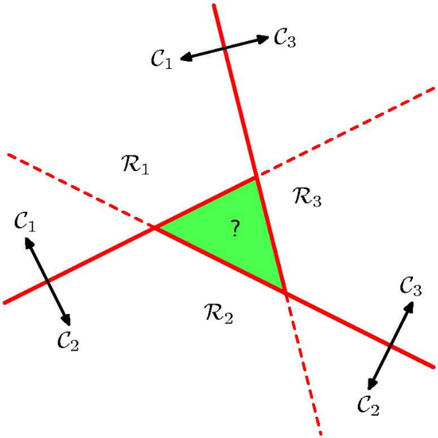
Classificação multiclasse

- **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 $\mathbf{y}_i$ consiste em um vetor de $K - 1$ zeros e um valor 1 na k-ésima posição associada à k-ésima classe.
- **Exemplo**: $\mathbf{y}_i = [1\ 0\ 0]^\top$, ou $\mathbf{y}_i = [0\ 1\ 0]^\top$, ou $\mathbf{y}_i = [0\ 0\ 1]^\top$.

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 $\mathbf{y}_i$ consiste em um vetor de $K-1$ zeros e um valor 1 na k-ésima posição associada à k-ésima classe.
- **Exemplo**: $\mathbf{y}_i = [1\ 0\ 0]^\top$, ou $\mathbf{y}_i = [0\ 1\ 0]^\top$, ou $\mathbf{y}_i = [0\ 0\ 1]^\top$.

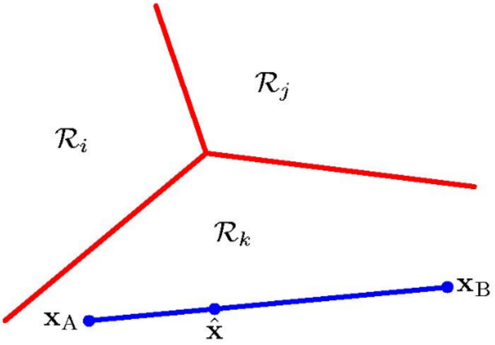
Discriminante linear

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

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

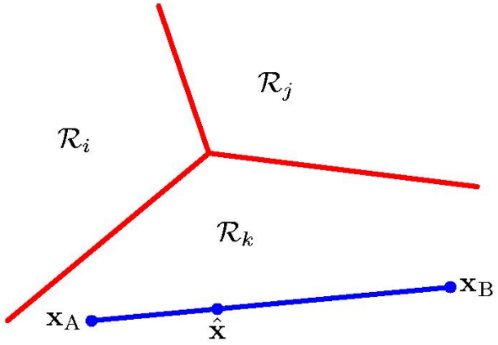
Classificação multiclasse

- 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:

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

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

Regressão multivariada

OLS para regressão multivariada (múltiplas saídas)

- **Função custo**:

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

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

- **Solução analítica**:

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

Regressão multivariada

Gradiente Descendente para múltiplas saídas

- Regra de atualização:

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

Gradiente Descendente Estocástico para múltiplas saídas

- Regra de atualização:

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

- Note que:

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

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

Classificação multiclasse

Regressão logística multiclasse

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

Classificação multiclasse

Regressão logística multiclasse

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

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

Classificação multiclasse

Regressão logística multiclasse

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

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

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

Classificação multiclasse

Regressão logística multiclasse

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

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

- Interpretação probabilística: $\hat{y}_{ik} = p(y_{ik}|\mathbf{x}_i, \mathbf{W}) \in [0, 1]$.
- 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:

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

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

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

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

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

Regressão logística multiclasse

- As derivadas em relação aos parâmetros são dadas por:

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

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

em que

$$\frac{\partial \hat{y}_{ik}}{\partial \mathbf{w}_k} = (\hat{y}_{ik} - \hat{y}_{ik}^2)\mathbf{x}_i, \quad j = k,$$
$$\frac{\partial \hat{y}_{ij}}{\partial \mathbf{w}_k} = -\hat{y}_{ij}\hat{y}_{ik}\mathbf{x}_i, \quad j \neq k,$$

ou seja,

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

Regressão logística multiclasse

- Substituindo na derivada original:

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

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

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

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

- Note que usamos acima o fato da soma dos elementos do vetor $\mathbf{y}_i$ ser igual a 1.

Regressão logística multiclasse

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

Gradiente Descendente para múltiplas saídas

- Regra de atualização:

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

Gradiente Descendente Estocástico para múltiplas saídas

- Regra de atualização:

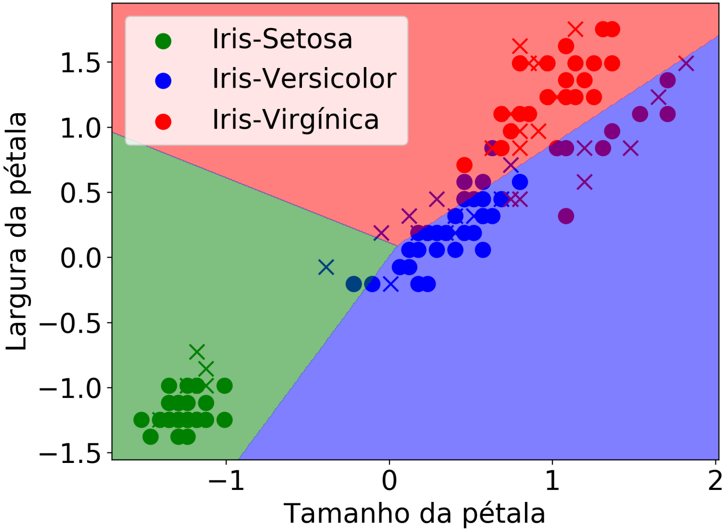
$$\mathbf{w}_k(t) = \mathbf{w}_k(t-1) + \alpha e_{ik}(t-1)\mathbf{x}_i$$

- Lembrando que na regressão logística multiclasse temos:

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

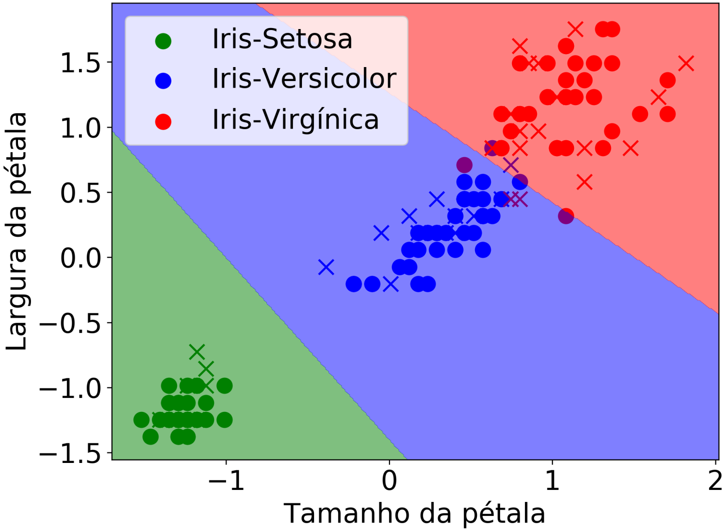
Classificação multiclasse

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



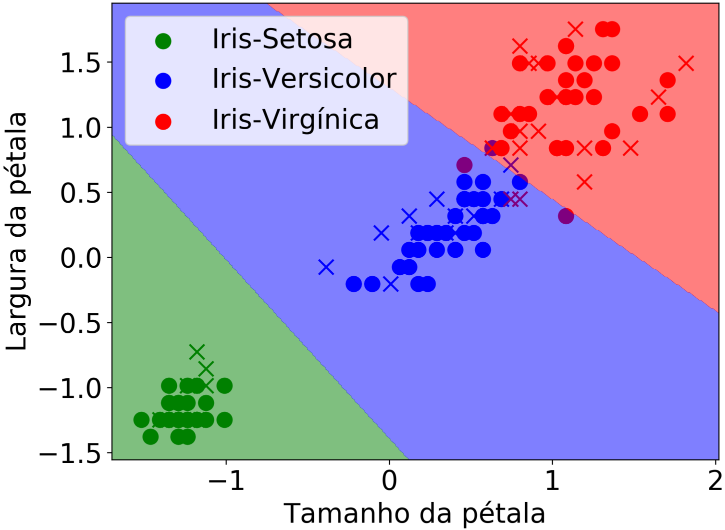
Classificação multiclasse

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



Classificação multiclasse

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



Agenda

Algoritmo IRLS

- Por causa da não-linearidade da função sigmoide, não há uma solução analítica para a regressão logística.

Algoritmo IRLS

- Por causa da não-linearidade da função sigmoide, não há uma solução analítica para a regressão logística.
- **IRLS (iterative reweighted least squares)**: Aplicação iterativa do algoritmo de Newton-Raphson:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - (\nabla\nabla\mathcal{J}(\mathbf{w}_{t-1}))^{-1}\nabla\mathcal{J}(\mathbf{w}_{t-1}),$$

em que $\mathcal{J}(\mathbf{w})$ é a função custo, $\nabla\mathcal{J}(\mathbf{w})$ é o seu gradiente em relação aos parâmetros $\mathbf{w}$ e $\nabla\nabla\mathcal{J}(\mathbf{w})$ é a sua Hessiana.

Algoritmo IRLS

- Computamos o gradiente e a Hessiana da função custo $\mathcal{J}(\mathbf{w})$:

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

$$\nabla \mathcal{J}(\mathbf{w}) = -\sum_{i=1}^{N} (y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)) \mathbf{x}_i = -\mathbf{X}^\top (\mathbf{y} - \sigma(\mathbf{X}\mathbf{w})),$$

$$\nabla \nabla \mathcal{J}(\mathbf{w}) = \sum_{i=1}^{N} \sigma(\mathbf{w}^\top \mathbf{x}_i)(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)) \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{X}^\top \mathbf{R} \mathbf{X},$$

$$\mathbf{R} = \mathrm{diag}(R_{11}, \ldots, R_{NN}),$$

$$R_{ii} = \sigma(\mathbf{w}^\top \mathbf{x}_i)(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)).$$

Algoritmo IRLS

- Os parâmetros $\mathbf{w}$ são atualizados via IRLS da seguinte maneira:

$$
\begin{aligned}
\mathbf{w}_t &= \mathbf{w}_{t-1} - (\nabla\nabla\mathcal{J}(\mathbf{w}_{t-1}))^{-1}\nabla\mathcal{J}(\mathbf{w}_{t-1}) \\
&= \mathbf{w}_{t-1} + (\mathbf{X}^\top\mathbf{R}_{t-1}\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}_{t-1})) \\
&= (\mathbf{X}^\top\mathbf{R}_{t-1}\mathbf{X})^{-1}\left[\mathbf{X}^\top\mathbf{R}_{t-1}\mathbf{X}\mathbf{w}_{t-1} + \mathbf{X}^\top(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}_{t-1}))\right] \\
&= (\mathbf{X}^\top\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{R}_{t-1}\mathbf{z}_{t-1},
\end{aligned}
$$

em que:

$$
\begin{aligned}
\mathbf{R}_{t-1} &= \mathrm{diag}([R_{t-1}]_{11}, \ldots, [R_{t-1}]_{NN}), \\
[R_{t-1}]_{ii} &= \sigma(\mathbf{w}_{t-1}^\top\mathbf{x}_i)(1 - \sigma(\mathbf{w}_{t-1}^\top\mathbf{x}_i)), \\
\mathbf{z}_{t-1} &= \mathbf{X}\mathbf{w}_{t-1} + \mathbf{R}_{t-1}^{-1}(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}_{t-1})).
\end{aligned}
$$

- Pode ser extrapolado para o caso multiclasse (Murphy, pgs. 252-254).

Algoritmo IRLS - Solução MAP

- Considerando uma priori $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$, temos:

$$\mathcal{J}_{\text{MAP}}(\mathbf{w}) = \mathcal{J}(\mathbf{w}) - \log \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).$$

- O gradiente e a Hessiana modificados serão dados por:

$$\begin{aligned}
\nabla \mathcal{J}_{\text{MAP}}(\mathbf{w}) &= \nabla \mathcal{J}(\mathbf{w}) + \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0), \\
\nabla\nabla \mathcal{J}_{\text{MAP}}(\mathbf{w}) &= \nabla\nabla \mathcal{J}(\mathbf{w}) + \mathbf{S}_0^{-1}.
\end{aligned}$$

- A solução iterativa MAP do IRLS é computada por:

$$\begin{aligned}
\mathbf{w}_t &= \mathbf{w}_{t-1} - (\nabla\nabla \mathcal{J}_{\text{MAP}}(\mathbf{w}_{t-1}))^{-1} \nabla \mathcal{J}_{\text{MAP}}(\mathbf{w}_{t-1}) \\
&= \mathbf{w}_{t-1} + \mathbf{A}^{-1}[\mathbf{X}^\top(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}_{t-1})) - \mathbf{S}_0^{-1}(\mathbf{w}_{t-1} - \mathbf{m}_0)],
\end{aligned}$$

em que:

$$\begin{aligned}
\mathbf{A} &= \mathbf{X}^\top \mathbf{R}_{t-1} \mathbf{X} + \mathbf{S}_0^{-1}, \\
\mathbf{R}_{t-1} &= \operatorname{diag}([R_{t-1}]_{11}, \ldots, [R_{t-1}]_{NN}), \\
[R_{t-1}]_{ii} &= \sigma(\mathbf{w}_{t-1}^\top \mathbf{x}_i)(1 - \sigma(\mathbf{w}_{t-1}^\top \mathbf{x}_i)).
\end{aligned}$$

Algoritmo IRLS - Solução MAP

Resumo do algoritmo

- Passo de estimação
  1. Defina a partir de conhecimentos/experimentos anteriores:
     - $\rightarrow$ os momentos da priori $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$;
     - $\rightarrow$ o valor inicial de $\mathbf{w}_0 \in \mathbb{R}^D$.
  2. A partir dos dados $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, repita até convergir:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \mathbf{A}^{-1}[\mathbf{X}^\top(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}_{t-1})) - \mathbf{S}_0^{-1}(\mathbf{w}_{t-1} - \mathbf{m}_0)],$$

em que:

$$\mathbf{A} = \mathbf{X}^\top \mathbf{R}_{t-1}\mathbf{X} + \mathbf{S}_0^{-1},$$
$$\mathbf{R}_{t-1} = \mathrm{diag}([R_{t-1}]_{11}, \ldots, [R_{t-1}]_{NN}),$$
$$[R_{t-1}]_{ii} = \sigma(\mathbf{w}_{t-1}^\top \mathbf{x}_i)(1 - \sigma(\mathbf{w}_{t-1}^\top \mathbf{x}_i)).$$

  3. Retorne os parâmetros estimados $\hat{\mathbf{w}}$.
- Passo de predição
  1. Dado um padrão $\mathbf{x}_*$, retorne a predição:

$$p(y_* = 1|\mathbf{x}_*) = \sigma(\hat{\mathbf{w}}^\top \mathbf{x}_*).$$

Agenda

Regressão logística Bayesiana

- Considere um modelo de regressão logística binária:

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \prod_{i=1}^{N} \sigma(\mathbf{w}^\top \mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))^{1-y_i}, \quad y_i \in \{0, 1\}, \\
&= \prod_{i=1}^{N} \sigma_i^{y_i} (1 - \sigma_i^{1-y_i}),
\end{aligned}
$$

em que $\sigma_i = \sigma(\mathbf{w}^\top \mathbf{x}_i) = \dfrac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}_i)}$.

- Considere uma priori Gaussiana para o vetor de parâmetros $\mathbf{w}$:

$$
p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).
$$

- Após a observação dos dados $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, como calcular a posteriori $p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$?

Regressão logística Bayesiana

- Considere um modelo de regressão logística binária:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{N} \sigma(\mathbf{w}^\top \mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))^{1-y_i}, \quad y_i \in \{0, 1\},$$
$$= \prod_{i=1}^{N} \sigma_i^{y_i} (1 - \sigma_i^{1-y_i}),$$

em que $\sigma_i = \sigma(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x}_i)}$.

- Considere uma priori Gaussiana para o vetor de parâmetros $\mathbf{w}$:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).$$

- Após a observação dos dados $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, como calcular a posteriori $p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$? Sem solução analítica!

Aproximação de Laplace

- **Ideia:** Aproximar uma distribuição complexa por uma Gaussiana.

Aproximação de Laplace

- **Ideia:** Aproximar uma distribuição complexa por uma Gaussiana.
- **Ideia:** A média será a moda da posteriori original e a matriz de covariância será a Hessiana do negativo da log-posteriori.

Aproximação de Laplace

- **Ideia:** Aproximar uma distribuição complexa por uma Gaussiana.
- **Ideia:** A média será a moda da posteriori original e a matriz de covariância será a Hessiana do negativo da log-posteriori. (calma, vamos justificar!)

Aproximação de Laplace

- **Ideia:** Aproximar uma distribuição complexa por uma Gaussiana.
- **Ideia:** A média será a moda da posteriori original e a matriz de covariância será a Hessiana do negativo da log-posteriori. (calma, vamos justificar!)
- Voltando à posteriori buscada:

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}) &= \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \\ \log p(\mathbf{w}|\mathcal{D}) &= \log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}) - \log p(\mathcal{D}), \\ \Psi(\mathbf{w}) &= \log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}). \end{aligned}$$

- A média da aproximação será a solução MAP $\hat{\mathbf{w}}$ usual:

$$\hat{\mathbf{w}} = \arg\max \Psi(\mathbf{w})$$

Aproximação de Laplace

Teorema de Taylor para aproximação de funções

- Uma função real k vezes diferenciável em a pode ser aproximada por um polinômio de Taylor de ordem k:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

- Por conveniência, podemos truncar a série em uma ordem $< k$.

Aproximação de Laplace

Teorema de Taylor para aproximação de funções

- Uma função real k vezes diferenciável em a pode ser aproximada por um polinômio de Taylor de ordem k:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

- Por conveniência, podemos truncar a série em uma ordem $< k$.

- Escolhemos uma expansão de Taylor de segunda ordem (quadrática) para aproximar $\Psi(\mathbf{w})$ em torno de $\hat{\mathbf{w}}$:

$$\begin{aligned}\Psi(\mathbf{w}) &\approx \Psi(\hat{\mathbf{w}}) + (\mathbf{w}-\hat{\mathbf{w}})^\top \Psi'(\hat{\mathbf{w}}) + \frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \Psi''(\hat{\mathbf{w}})(\mathbf{w}-\hat{\mathbf{w}}) \\ &\approx \Psi(\hat{\mathbf{w}}) + \frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \Psi''(\hat{\mathbf{w}})(\mathbf{w}-\hat{\mathbf{w}}),\end{aligned}$$

em que $\Psi'(\hat{\mathbf{w}}) = 0$, pois $\hat{\mathbf{w}}$ é um máximo de $\Psi(\mathbf{w})$.

Aproximação de Laplace

- Podemos reescrever a aproximação:

$$\Psi(\mathbf{w}) \approx \Psi(\hat{\mathbf{w}}) - \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{H}(\mathbf{w} - \hat{\mathbf{w}}),$$

em que $\mathbf{H} = -\Psi''(\hat{\mathbf{w}}) = -\nabla\nabla\Psi(\hat{\mathbf{w}})$

- Aplicamos uma exponencial em ambos os lados:

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}) &\overset{\propto}{\sim} \exp(\Psi(\mathbf{w})) \\ &\overset{\propto}{\sim} \exp(\Psi(\hat{\mathbf{w}})) \exp\left(-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{H}(\mathbf{w} - \hat{\mathbf{w}})\right), \\ &\approx \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1}). \end{aligned}$$

Aproximação de Laplace

- Podemos também aproximar a evidência:

$$
\begin{aligned}
\log p(\mathcal{D}) &= \log \int p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}) \mathrm{d}\mathbf{w} \\
&\approx \log \int \exp(\Psi(\mathbf{w})) \mathrm{d}\mathbf{w} \\
&\approx \Psi(\hat{\mathbf{w}}) + \log \int \exp\left(-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{H}(\mathbf{w} - \hat{\mathbf{w}})\right) \mathrm{d}\mathbf{w} \\
&\approx \log p(\mathcal{D}|\hat{\mathbf{w}}) + \log p(\hat{\mathbf{w}}) + \log \frac{(2\pi)^{M/2}}{|\mathbf{H}|^{1/2}} \\
&\approx \log p(\mathcal{D}|\hat{\mathbf{w}}) + \log p(\hat{\mathbf{w}}) + \frac{M}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{H}|.
\end{aligned}
$$

Bayesian Information Criterion (BIC)

- Repetimos a evidência aproximada abaixo:

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\hat{\mathbf{w}}) + \log p(\hat{\mathbf{w}}) + \frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{H}|.$$

- O índice BIC para comparação de modelos pode ser obtido após algumas considerações:
  - $\rightarrow$ Ignorar o termo constante;
  - $\rightarrow$ Uma priori uniforme para os parâmetros, o que torna $\hat{\mathbf{w}} = \mathbf{w}_{\text{ML}}$;
  - $\rightarrow$ Seja $\mathbf{H} = \sum_{i=1}^{N} \mathbf{H}_i$, em que $\mathbf{H}_i = \nabla\nabla \log p(\mathcal{D}_i|\mathbf{w})$. Aproximando $\mathbf{H}_i = \hat{\mathbf{H}}$, podemos aproximar o termo $\log|\mathbf{H}|$:

$$\log|\mathbf{H}| \approx \log|N\hat{\mathbf{H}}| = \log(N^D|\hat{\mathbf{H}}|) = D\log N + \log|\hat{\mathbf{H}}|,$$

    sendo D a dimensão de $\mathbf{w}$. O termo $\log|\hat{\mathbf{H}}|$ não depende de N e será dominado para N grande, podendo ser desprezado.

- A partir dos pontos acima o índice BIC é dado por:

$$\text{BIC} = \log p(\mathcal{D}|\mathbf{w}_{\text{ML}}) - \frac{D}{2}\log N.$$

Bayesian Information Criterion (BIC)

- Índice BIC para regressão linear/logística:

$$\mathrm{BIC} = \log p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - \frac{D}{2} \log N.$$

- De maneira mais geral, podemos considerar os graus de liberdade $\mathrm{dof}(\mathbf{w})$ do modelo:

$$\mathrm{BIC} = \log p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - \frac{\mathrm{dof}(\mathbf{w})}{2} \log N.$$

- Outro índice popular é o **Akaike Information Criterion (AIC)**:

$$\mathrm{AIC} = \log p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - \mathrm{dof}(\mathbf{w}).$$

Regressão Logística Bayesiana

- Dada uma entrada $\mathbf{x}_*$, buscamos uma distribuição preditiva:

$$
\begin{aligned}
p(y_* = 1|\mathbf{x}_*, \mathcal{D}) &= \int p(y_*|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|\mathcal{D})\mathrm{d}\mathbf{w} \\
&\approx \int \sigma(\mathbf{w}^\top \mathbf{x}_*)\mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1})\mathrm{d}\mathbf{w}, \\
p(y_* = 0|\mathbf{x}_*, \mathcal{D}) &= 1 - p(y_* = 1|\mathbf{x}_*, \mathcal{D}).
\end{aligned}
$$

- A matriz $\mathbf{H} = -\nabla\nabla\Psi(\hat{\mathbf{w}})$ será dada por:

$$
\begin{aligned}
&\mathbf{H} = -\nabla\nabla[\log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w})]|_{\hat{\mathbf{w}}} = \mathbf{X}^\top \hat{\mathbf{R}}\mathbf{X} + \mathbf{S}_0^{-1}, \\
&\hat{\mathbf{R}} = \mathrm{diag}(\hat{R}_{11}, \dots, \hat{R}_{NN}), \quad \hat{R}_{ii} = \sigma(\hat{\mathbf{w}}^\top \mathbf{x}_i)(1 - \sigma(\hat{\mathbf{w}}^\top \mathbf{x}_i)).
\end{aligned}
$$

- A distribuição preditiva não é analítica e precisa ser aproximada:
  - $\rightarrow$ Aproximação de Monte Carlo;
  - $\rightarrow$ Aproximação probit.

Regressão Logística Bayesiana - Monte Carlo

- Seguindo uma aproximação de **Monte Carlo**:

$$
\begin{aligned}
p(y_* = 1|\mathbf{x}_*, \mathcal{D}) &\approx \int \sigma(\mathbf{w}^\top \mathbf{x}_*)\mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1})\mathrm{d}\mathbf{w}, \\
&= \frac{1}{S}\sum_{s=1}^{S} \sigma(\mathbf{w}_s^\top \mathbf{x}_*), \\
\mathbf{w}_s &\sim \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1}).
\end{aligned}
$$

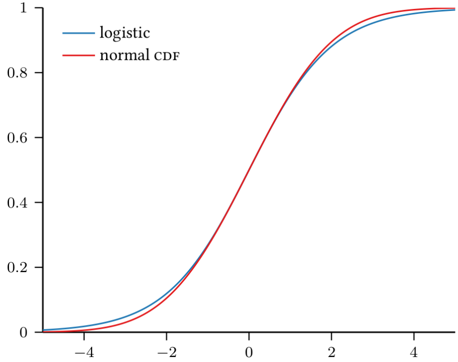
- Gerar amostras independentes de uma Gaussiana multivariada pode ser feito diretamente via pacotes estatísticos.
- Podemos reaproveitar as amostras $\mathbf{w}_s|_{s=1}^{S}$ para diferentes entradas.

Regressão Logística Bayesiana - aproximação probit

- A função logística $\sigma(z)$ pode ser aproximada pela função cumulativa (CDF) da distribuição Gaussiana normalizada:

$$\sigma(z) \approx \Phi(\lambda z), \quad \lambda = \sqrt{\pi/8},$$

$$\Phi(z) \triangleq \int_{-\infty}^{z} \mathcal{N}(z|0,1)\mathrm{d}z \quad \text{(função probit).}$$



Regressão Logística Bayesiana - aproximação probit

- Substituímos a função logística pela função probit:

$$p(y_* = 1|\mathbf{x}_*, \mathcal{D}) \approx \int \Phi(\lambda \mathbf{w}^\top \mathbf{x}_*) \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1}) \mathrm{d}\mathbf{w}$$

- Note que o termo $\Phi(\lambda \mathbf{w}^\top \mathbf{x}_*)$ na verdade depende de um escalar $a = \mathbf{w}^\top \mathbf{x}_*$. Por isso, a integral se torna unidimensional:

$$p(y_* = 1|\mathbf{x}_*, \mathcal{D}) \approx \int \Phi(\lambda a) \mathcal{N}(a|\mu_a, \sigma_a^2) \mathrm{d}a,$$

$$\mu_a = \hat{\mathbf{w}}^\top \mathbf{x}_*,$$

$$\sigma_a^2 = \mathbf{x}_*^\top \mathbf{H}^{-1} \mathbf{x}_*.$$

- Essa última integral possui a solução analítica abaixo:

$$p(y_* = 1|\mathbf{x}_*, \mathcal{D}) \approx \Phi\left(\frac{\mu_a}{(\lambda^{-2} + \sigma_a^2)^{1/2}}\right) = \sigma((1 + \pi\sigma_a^2/8)^{1/2}\mu_a).$$

Regressão Logística Bayesiana

Resumo do algoritmo

- Passo de estimação
  1. Defina a partir de conhecimentos/experimentos anteriores:
     $\rightarrow$ os momentos da priori $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$;
  2. A partir dos dados $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, encontre a solução MAP para $\hat{\mathbf{w}}$ (e.g. via algoritmo IRLS).
  3. Aproxime a posteriori de $\mathbf{w}$:
     $$p(\mathbf{w}|\mathcal{D}) \approx \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1}),$$
     em que:
     $$\mathbf{H} = \mathbf{X}^\top \hat{\mathbf{R}} \mathbf{X} + \mathbf{S}_0^{-1},$$
     $$\hat{\mathbf{R}} = \operatorname{diag}(\hat{R}_{11}, \ldots, \hat{R}_{NN}),$$
     $$\hat{R}_{ii} = \sigma(\hat{\mathbf{w}}^\top \mathbf{x}_i)(1 - \sigma(\hat{\mathbf{w}}^\top \mathbf{x}_i)).$$
  4. Retorne a posteriori aproximada $p(\mathbf{w}|\mathcal{D})$ dos parâmetros.

Regressão Logística Bayesiana

Resumo do algoritmo

- Passo de predição
  1. Dado um padrão $\mathbf{x}_*$, retorne a distribuição preditiva:
     - $\rightarrow$ Via aproximação de Monte Carlo:

$$p(y_* = 1|\mathbf{x}_*) \approx \frac{1}{S}\sum_{s=1}^{S}\sigma(\mathbf{w}_s^\top \mathbf{x}_*),$$

$$\mathbf{w}_s \sim \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{H}^{-1}).$$

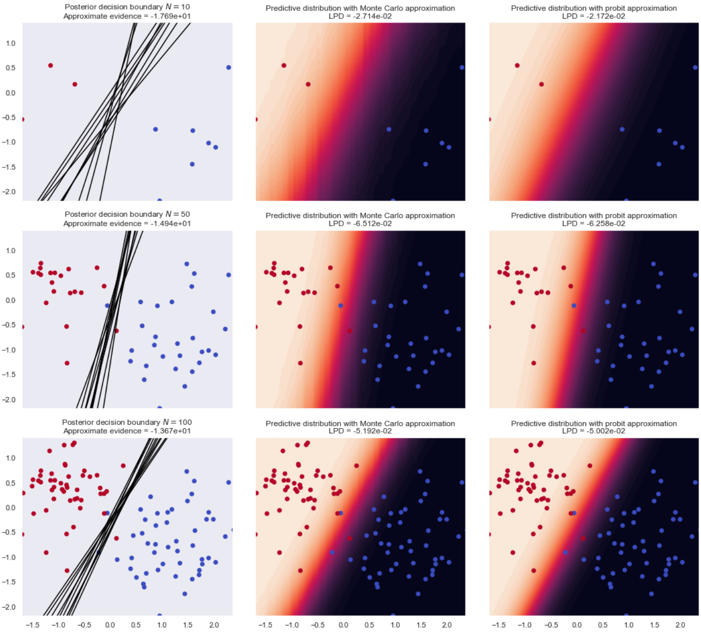
     - $\rightarrow$ Via aproximação probit:

$$p(y_* = 1|\mathbf{x}_*) \approx \sigma((1 + \pi\sigma_a^2/8)^{1/2}\mu_a),$$

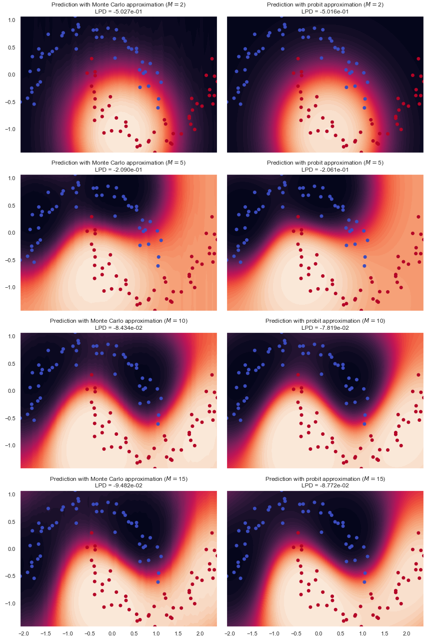
$$\mu_a = \hat{\mathbf{w}}^\top \mathbf{x}_*,$$

$$\sigma_a^2 = \mathbf{x}_*^\top \mathbf{H}^{-1}\mathbf{x}_*.$$

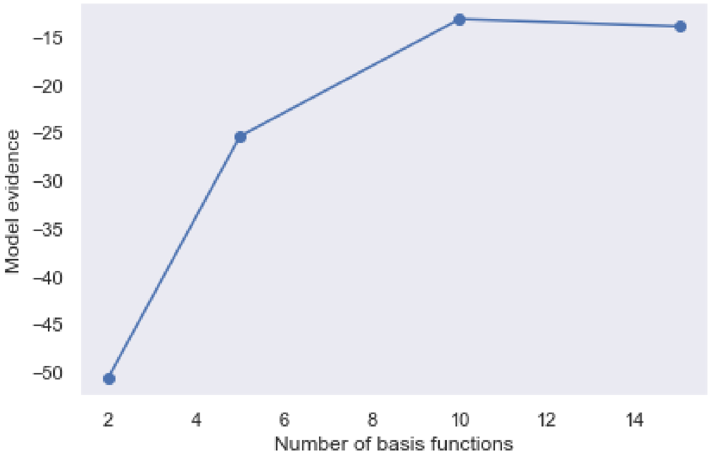
Regressão Logística Bayesiana



Regressão Logística RBF Bayesiana



Regressão Logística RBF Bayesiana



Agenda

Tópicos adicionais

- Representação de atributos categóricos via one hot encoding.
  - $\rightarrow$ **Exemplo**: Atributo "gênero de filme" (ação, drama ou comédia): $\mathbf{x}_i = [1\ 0\ 0]^\top$, ou $\mathbf{x}_i = [0\ 1\ 0]^\top$, ou $\mathbf{x}_i = [0\ 0\ 1]^\top$.
- Generalized linear models (GLMs).
- Regressão ordinal.

Agenda

Referências bibliográficas

- **Cap. 8** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 4** - BISHOP, Christopher M. **Pattern recognition and machine learning**, 2006.