La Regression logistique

February 12, 2024

La régression logistique est une technique de modélisation statistique utilisée pour prédire la probabilité qu'une variable descriptive binaire prenne l'une des deux valeurs possibles (0 ou 1), on peut noter Y cette variable (elle appartient à $\{0,1\}^n$) en fonction d'un ensemble de variables explicatives que l'on note $X = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{n\times p}$, avec n observations et p variables. C'est une méthode couramment utilisée en apprentissage automatique et en statistiques.

La régression logistique a une interprétation probabiliste, elle permet de modéliser $P(Y = 1|\boldsymbol{x})$,où $\boldsymbol{x} \in \mathbb{R}^p$.

En utilisant la loi de Bayes et le fait que :

$$P(\boldsymbol{x}) = P(\boldsymbol{x}|Y = 1)P(Y = 1) + P(\boldsymbol{x}|Y = 0)P(Y = 0)$$

nous avons :

$$\begin{aligned} P(Y = 1|\boldsymbol{x}) &= \frac{P(\boldsymbol{x}|Y = 1)P(Y = 1)}{P(\boldsymbol{x}|Y = 1)P(Y = 1) + P(\boldsymbol{x}|Y = 0)P(Y = 0)} \\ &= \frac{1}{1 + \frac{P(\boldsymbol{x}|Y=0)P(Y=0)}{P(\boldsymbol{x}|Y=1)P(Y=1)}} \\ &= \frac{1}{1 + \frac{P(Y=0|\boldsymbol{x})}{P(Y=1|\boldsymbol{x})}} \end{aligned}$$

On note $f(\boldsymbol{x}) := \log\left(\frac{P(Y=1|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right)$.

On a ainsi $P(Y = 1|\boldsymbol{x}) =: \sigma(f(\boldsymbol{x}))$ avec $\sigma(z) = \frac{1}{1+e^{-z}}$.

La fonction σ, appelée fonction logistique, satisfait les propriétés suivantes :

$$\begin{aligned} \sigma(-z) &= 1 - \sigma(z) \\ \frac{d\sigma(z)}{dz} &= \sigma(z)\sigma(-z). \end{aligned}$$

L'intérêt de la fonction logistique réside dans sa capacité à transformer une fonction f à valeurs dans $\mathbb{R}$ en une probabilité comprise entre 0 et 1.

La régression logistique revient en fait à supposer que f est linéaire de la forme $f : \boldsymbol{x} \mapsto \boldsymbol{\theta}^\top \boldsymbol{x}$ avec $\boldsymbol{\theta} \in \mathbb{R}^p$.

Sous cette hypothèse, la règle de classification est simplement :

$$\begin{cases} \text{si } \boldsymbol{\theta}^\top \boldsymbol{x} \leq 0, \text{ on étiquette 0 au point } \boldsymbol{x} \\ \text{si } \boldsymbol{\theta}^\top \boldsymbol{x} > 0, \text{ on étiquette 1 au point } \boldsymbol{x} \end{cases}$$

On obtient donc :

$$\begin{aligned} P(Y = 1|\boldsymbol{x}) &= \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) \\ P(Y = 0|\boldsymbol{x}) &= 1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) = \sigma(-\boldsymbol{\theta}^\top \boldsymbol{x}) \end{aligned}$$

Le but maintenant est d'estimer $\boldsymbol{\theta}$. Nous avons $(x_i, y_i)_{1 \leq i \leq n}$ où $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$ qui constitue un échantillon de taille n. . On a alors :

$$P(y = y_i|\boldsymbol{x} = x_i) = \sigma(\boldsymbol{\theta}^\top x_i)^{y_i} \sigma(-\boldsymbol{\theta}^\top x_i)^{1-y_i}.$$

La log-vraisemblance s'exprime alors de cette manière :

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log \left(\sigma(\boldsymbol{\theta}^\top x_i)^{y_i} \sigma(-\boldsymbol{\theta}^\top x_i)^{1-y_i} \right) \\ &= \sum_{i=1}^{n} l(\boldsymbol{\theta}^\top x_i, y_i). \end{aligned}$$

où l est définie comme étant la fonction de perte logistique. Il faut ensuite avoir recours à des algorithmes itératifs (descente de gradient, méthode de Newton,...) pour trouver $\hat{\boldsymbol{\theta}}$.

On peut passer du cadre binaire au cadre multi-classes avec K classes, par exemple, c'est-à-dire en ayant Y appartenant à $[\![1, K]\!]^n$. De nouveau, on modélise les probabilités conditionnelles des classes, ou plutôt leur log-ratio, par des quantités linéaires : $\log \left(\frac{P(Y=k|\boldsymbol{x})}{P(Y=K|\boldsymbol{x})} \right) = \theta_k^\top \boldsymbol{x}_i$ pour $k \in [\![1, K-1]\!]$ et $\theta_k \in \mathbb{R}^p$

On a alors pour paramètre globale: $\theta \in [\theta_1, \ldots, \theta_K] \in \mathbb{R}^{p \times K}$ et pour $k \in [\![1, K]\!]$:

$$\begin{aligned} P(y = k|\boldsymbol{x}) &= \frac{\exp(\langle \boldsymbol{\theta}_k, \boldsymbol{x} \rangle)}{\sum_{l=1}^{K} \exp(\langle \boldsymbol{\theta}_l, \boldsymbol{x} \rangle)}, \\ &= \frac{\exp(\boldsymbol{\theta}_k^T \boldsymbol{x})}{\sum_{l=1}^{K} \exp(\boldsymbol{\theta}_l^T \boldsymbol{x})} \end{aligned}$$

On peut écrire cette égalité sous forme vectorielle en utilisant la notation softmax, on a alors : $(P(y = k|\boldsymbol{x}))_{k=1,\ldots,K} = softmax(\theta_1^T \boldsymbol{x}, .., \theta_K^T \boldsymbol{x})$

Pour la régression softmax, la log-vraisemblance peut être exprimée comme suit :

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(y_i = k) \log \left(\frac{e^{\theta_k^\top x_i}}{\sum_{l=1}^{K} e^{\theta_l^\top x_i}} \right)$$

On utilise ensuite des méthodes algortihmiques pour trouver $\hat{\boldsymbol{\theta}}$.