Universidad de los Andes > FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

FACULTAD DE INGENIERIA Y CIENCIAS APLICADAS

Artificial Intelligence

Introduction

Machine Learning

Carla Vairetti

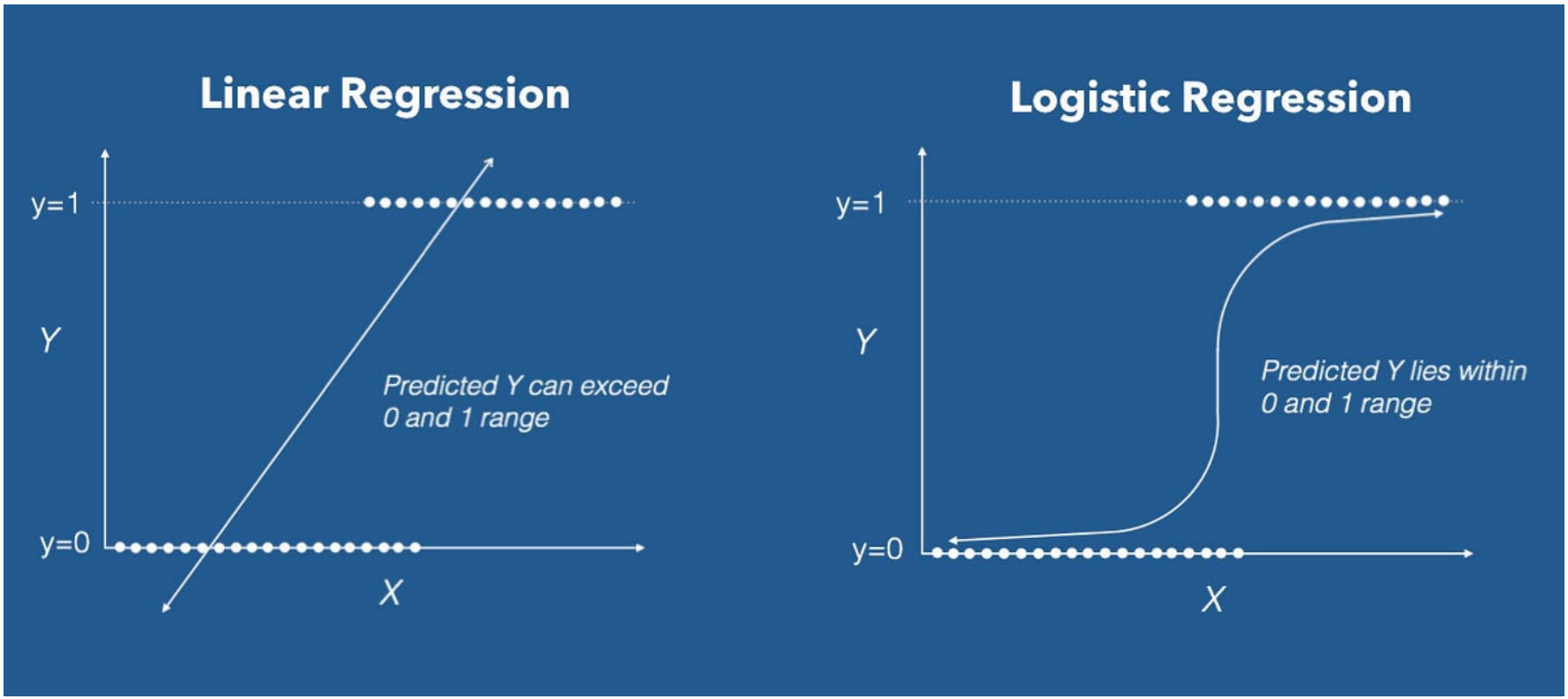
cvairetti@uandes.cl - carla.vairetti@miuandes.cl

ING/UAndes

06-08-2024

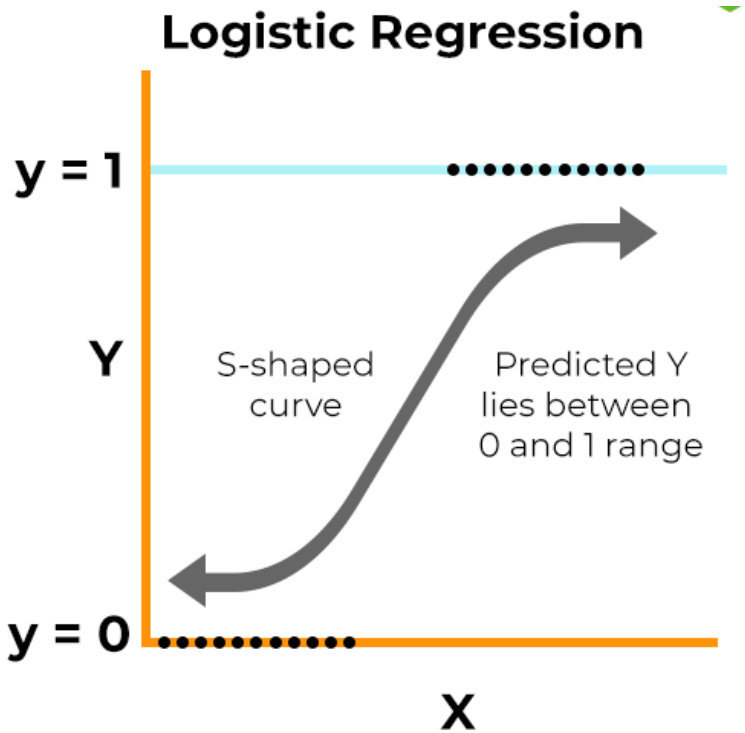
Machine Learning

Logistic Regression



Logistic Regression

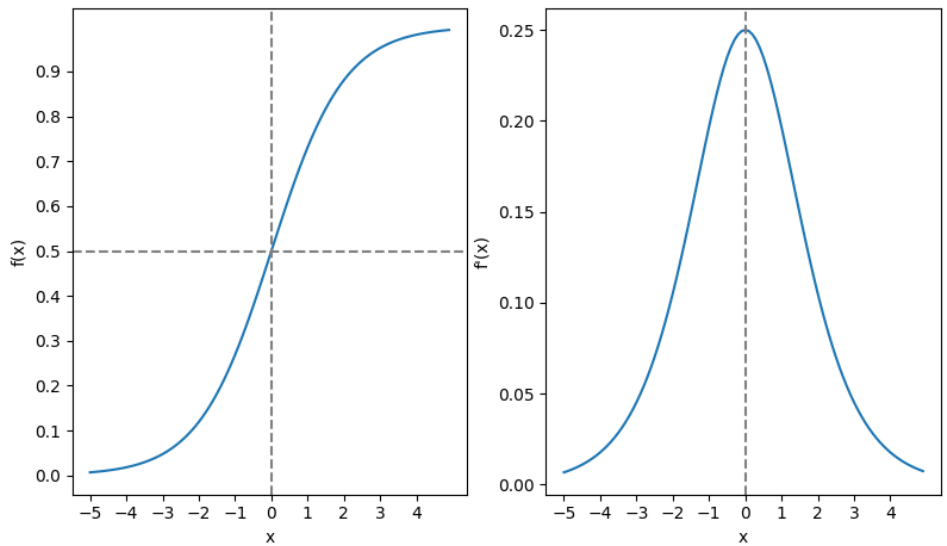
- This is an extension of the linear regression model to handle classification problems, where the output y is associated with discrete values. To facilitate understanding, we will assume a binary classification problem ($y \in \{0, 1\}$).



To address the above problem, we apply a function that restricts the output of the model $\mathcal{F}_\Phi(\mathrm{x})$ a $[0, 1]$. This function is the *logistic function* or *sigmoid function* defined by:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

whose derivative is $f'(x) = f(x)(1 - f(x))$ (it's a good exercise to derive this).



- Thus, the predictor function $\mathcal{F}_\Phi(\cdot)$ remains defined by Eq. 2.

$$\mathcal{F}_\Phi(\mathrm{x}) = f(\Phi^T \mathrm{x}) \tag{2}$$

$$= \frac{1}{1 + e^{-\Phi^T \mathrm{x}}} \tag{3}$$

Given the previous definitions, the next step is to determine how to find Φ. To do this, we need to recall some basic concepts from Calculus and Probability. Thus, we will interpret the classification model from a probabilistic perspective, so that Φ can be estimated using an approach known as **Maximum Likelihood Estimation**.

We assume that:

$$P(y = 1|\mathrm{x}; \Phi) = \mathcal{F}_\Phi(\mathrm{x}) \tag{4}$$
$$P(y = 0|\mathrm{x}; \Phi) = 1 - \mathcal{F}_\Phi(\mathrm{x}) \tag{5}$$

Both previous expressions can be written in a compact form as:

$$p(y|\mathrm{x}; \Phi) = (\mathcal{F}_\Phi(\mathrm{x}))^y (1 - \mathcal{F}_\Phi(\mathrm{x}))^{1-y} \tag{6}$$

Assuming that the N training data were generated independently, we can write the *likelihood* of the model $\mathbf{L}(\Phi)$:

$$\mathbf{L}(\Phi) = p(\vec{y}|\mathrm{X}; \Phi) \tag{7}$$
$$= \Pi_{i=1}^N p(y_i|\mathrm{x}_i; \Phi) \tag{8}$$
$$= \Pi_{i=1}^N (\mathcal{F}_\Phi(\mathrm{x}_i))^{y_i} (1 - \mathcal{F}_\Phi(\mathrm{x}_i))^{1-y_i} \tag{9}$$

Likelihood L(Φ) of a parametric model defined by Φ is the conditional probability of generating the correct outputs under Φ.
To facilitate the calculations, it is advisable to work with the logarithm of the likelihood $\log(\mathcal{F}\Phi(\cdot))$. Thus, for the case of the logistic regression model, we have:

$$\log \mathbf{L}(\Phi) = \log \Pi_{i=1}^{N} (\mathcal{F}_{\Phi}(\mathrm{x}_i))^{y_i} (1 - \mathcal{F}_{\Phi}(\mathrm{x}_i))^{1-y_i} \tag{10}$$

$$= \sum_{i=1}^{N} [\log(\mathcal{F}_{\Phi}(\mathrm{x}_i))^{y_i} + \log(1 - \mathcal{F}_{\Phi}(\mathrm{x}_i))^{1-y_i}] \tag{11}$$

$$= \sum_{i=1}^{N} [y_i \log(\mathcal{F}_{\Phi}(\mathrm{x}_i)) + (1 - y_i) \log(1 - \mathcal{F}_{\Phi}(\mathrm{x}_i))] \tag{12}$$

Given Eq 12, the goal is to set Φ by maximizing $\log \mathbf{L}(\Phi)$. This task is known as **Maximum Likelihood Estimation** (MLE).

One of the most efficient ways to solve the MLE problem is through an iterative process guided by gradient ascent. This involves iteratively moving Φ according to the gradient of $\log \mathbf{L}(\Phi)$, using the following update function:

$$\Phi = \Phi + \alpha \nabla_{\Phi} \log \mathbf{L}(\Phi) \tag{13}$$

We will perform an analysis for the input (x_i, y_i).

Let's remember that $\nabla_{\Phi}(F) = \left[\frac{\partial F}{\partial \phi_1}, \cdots \frac{\partial F}{\phi_d}\right]$.

Let's see how to estimate $\frac{\partial}{\partial \phi_j} \log \mathbf{L}(\Phi)$:

$$
\begin{aligned}
\frac{\partial}{\partial \phi_j} \log \mathbf{L}(\Phi) &= \frac{\partial}{\partial \phi_j} y_i \log(\mathcal{F}_\Phi(\mathrm{x}_i)) + (1 - y_i) \log(1 - \mathcal{F}_\Phi(\mathrm{x}_i)) \\
&= \frac{\partial}{\partial \phi_j} y_i \log(\mathcal{F}_\Phi(\mathrm{x}_i)) + \frac{\partial}{\partial \phi_j} (1 - y_i) \log(1 - \mathcal{F}_\Phi(\mathrm{x}_i)) \\
&= \frac{\partial}{\partial \phi_j} y_i \log(f(\Phi^T \mathrm{x}_i)) + \frac{\partial}{\partial \phi_j} (1 - y_i) \log(1 - f(\Phi^T \mathrm{x}_i))
\end{aligned} \tag{14}
$$

$$\begin{aligned}
\frac{\partial}{\partial \phi_j} y_i \log(f(\Phi^T \mathrm{x}_i)) &= y_i \frac{1}{f(\Phi^T \mathrm{x}_i)} f(\Phi^T \mathrm{x}_i)(1 - f(\Phi^T \mathrm{x}_i)) x_{i,j} \\
&= y_i x_{i,j} - y_i f(\Phi^T \mathrm{x}_i) x_{i,j} \qquad (15) \\
\frac{\partial}{\partial \phi_j} (1 - y_i) \log(1 - f(\Phi^T \mathrm{x}_i)) &= (1 - y_i) \frac{1}{1 - f(\Phi^T \mathrm{x}_i)} (-f(\Phi^T \mathrm{x}_i))(1 - f(\Phi^T \mathrm{x}_i)) x_{i,j} \\
&= (y_i - 1) f(\Phi^T \mathrm{x}_i) x_{i,j} \\
&= y_i (f(\Phi^T \mathrm{x}_i)) x_{i,j} - f(\Phi^T \mathrm{x}_i) x_{i,j} \qquad (16)
\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \phi_i} \log \mathbf{L}(\Phi) &= y_i x_{i,j} - f(\Phi^T \mathrm{x}_i) x_{i,j} \\ &= (y_i - f(\Phi^T \mathrm{x}_i)) x_{i,j} \\ &= (y_i - \mathcal{F}_\Phi(\mathrm{x}_i)) x_{i,j} \end{aligned} \tag{17}$$

$$\phi_j = \phi_j + \alpha (y_i - \mathcal{F}_\Phi(\mathrm{x}_i)) x_{i,j} \tag{18}$$

$$\phi_j = \phi_j + \alpha \frac{1}{N} \left(\sum_{i=1}^{N} (y_i - \mathcal{F}_\Phi(\mathrm{x}_i)) x_{i,j} \right) \tag{19}$$

We have seen two linear models, one for the case of regression, which we call linear regression, and the other for the case of classification. This latter model is known as logistic regression. Both cases can be interpreted in probabilistic terms as follows:

- **Regression**

$$y|x;\Phi \sim \mathcal{N}(\mu, \sigma^2) \tag{20}$$

- **Classification**

$$y|x;\Phi \sim \text{Bernoulli}(\theta) \tag{21}$$

Given the above, it is possible to define a **family of distributions** that generalizes the previously seen linear models. This family is known as the Exponential Family and is defined as:

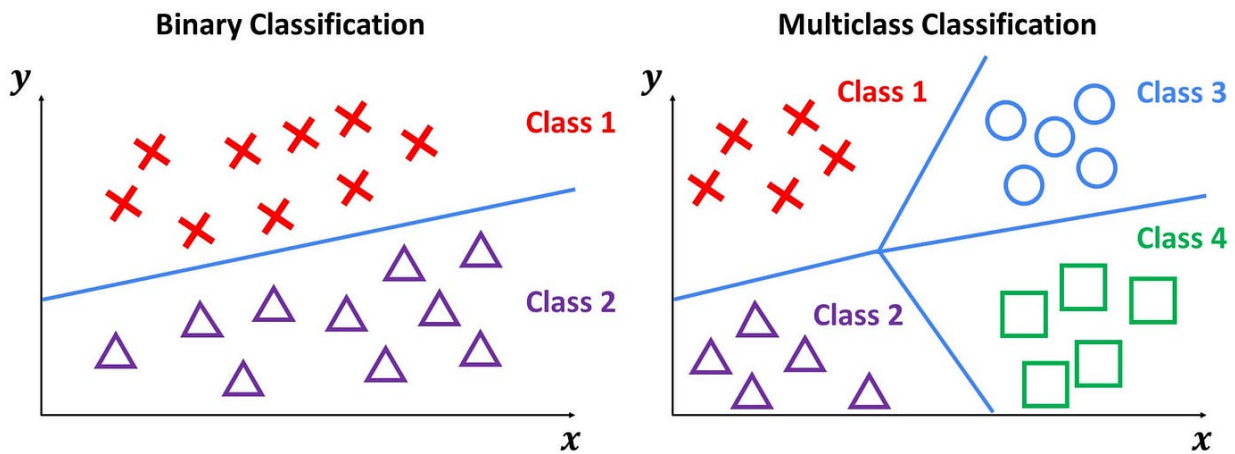
$$p(y;\eta) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right), \tag{22}$$

where :

- η is the *natural* parameter associated with the input data.
- $T(y)$ is a statistic about y.
- $a(\eta)$ represents a normalization factor.

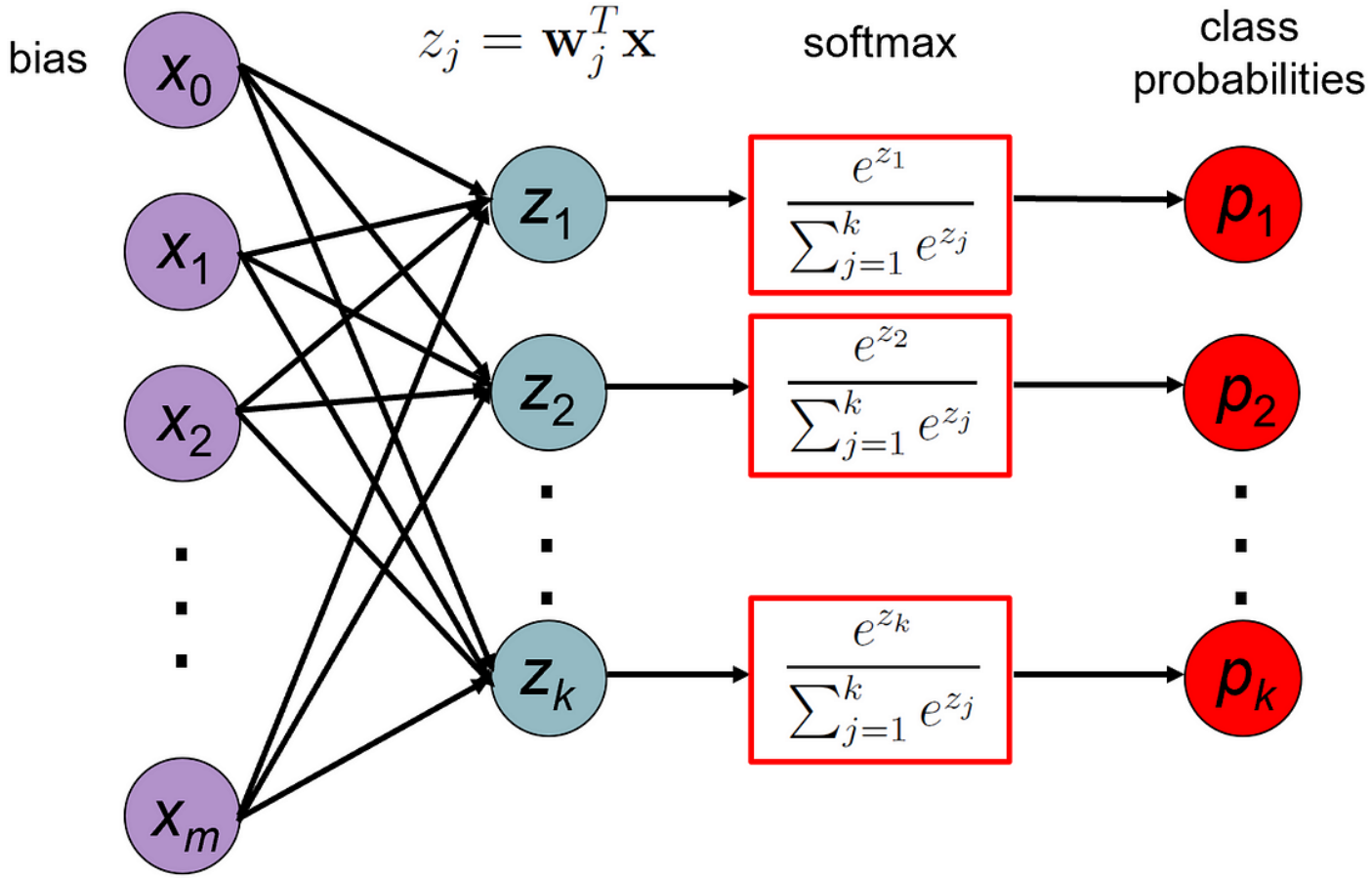
Thus, by setting T, a and b, a family of distributions is defined, parameterized by η.

Given the generalization of a linear model (GLM), we will extend the logistic regression model to the case of inferring multiple classes.
In a multiclass problem where $y \in \{1, 2, \cdots, k\}$, where k being the number of classes to infer.
For example, we can consider email classification, where an email can be classified as *spam*, *not spam*, *work email*, or *personal email*.



Softmax Training

Our goal is to estimate Φ in order to maximize $\mathbb{E}[T(y)|x;\Phi]$. To achieve this, we will follow the **Maximum Likelihood Estimation** (MLE) strategy using gradient ascent.



facultad de ingenieria y ciencias aplicadas

Artificial Intelligence

Introduction

Machine Learning

Carla Vairetti
cvairetti@uandes.cl - carla.vairetti@miuandes.cl

ING/UAndes

06-08-2024