Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

GLM and GAM

When we are modeling a ranomd variable Y from other random variables $X_1, ..., X_p$ in general we assume:

$$\boxed{Y = f(X_1, ..., X_p) + \epsilon}$$

with $\epsilon \sim D(0, \sigma^2)$ the error random variable which is independent of $X_1, ..., X_p$.

$\Rightarrow$

$$\mathbb{E}[Y|X_1, ..., X_p] = \mathbb{E}[f(X_1, ..., X_p) + \epsilon|X_1, ..., X_p]$$

$$= \mathbb{E}[f(X_1, ..., X_p)|X_1, ..., X_p] + \mathbb{E}[\epsilon|X_1, ..., X_p]$$

$$= f(X_1, ..., X_p) + \mathbb{E}[\epsilon] = f(X_1, ..., X_p)$$

$\Rightarrow$

$$\boxed{\mathbb{E}[Y|X_1, ..., X_p] = f(X_1, ..., X_p)}$$

The simplest function we can define is:

$$f(X_1, ..., X_p) = \alpha_0 + \alpha_1 \cdot X_1 + ... + \alpha_p \cdot X_p$$

$\Rightarrow$

$$\boxed{\mathbb{E}[Y|X_1, ..., X_p] = \alpha_0 + \alpha_1 \cdot X_1 + ... + \alpha_p \cdot X_p}$$

which is the known **Linear Regression**.

Linear regression is a **GLM** (Generalized Linear Model).

GLM: Generalized Linear Model

GLM is defined by three components:

- Probability distribution.
- Linear predictor: $h(X_1, ..., X_p) = \alpha_0 + \alpha_1 \cdot X_1 + ... + \alpha_p \cdot X_p$.
- Link function: $g(\cdot)$

All the GLMs satisfies the next equation:

$$g\left(\mathbb{E}[Y|X_1, ..., X_p]\right) = h(X_1, ..., X_p) = \alpha_0 + \alpha_1 \cdot X_1 + ... + \alpha_p \cdot X_p$$

$\Rightarrow$

$$\boxed{g\left(\mathbb{E}[Y|X_1, ..., X_p]\right) = \alpha_0 + \alpha_1 \cdot X_1 + ... + \alpha_p \cdot X_p}$$

$\Rightarrow$

$$\boxed{\mathbb{E}[Y|X_1, ..., X_p] = g^{-1}\left(\alpha_0 + \alpha_1 \cdot X_1 + ... + \alpha_p \cdot X_p\right)}$$

We can see the inverse of $g(\cdot)$ as a **activation function**.

Example: We can use $g(x) = logit(x) = \sigma^{-1}(x) = ln\left(\frac{x}{1-x}\right)$ and we have:

$$\mathbb{E}[Y|X_1, ..., X_p] = \sigma\left(\alpha_0 + \alpha_1 \cdot X_1 + ... + \alpha_p \cdot X_p\right)$$

You can see that is the **Logistic Regression**. In other word **Logistic Regression** is a **GLM**.

Examples of Link Functions:

Name	**Function**
Identity	$g(x) = x$
Log	$g(x) = ln(x)$
Logit	$g(x) = ln\left(\frac{x}{1-x}\right)$
Probit	$g(x) = \Phi^{-1}(x)$

With $\Phi(\cdot)$ the cumulative distribution function of the normal.

GAM: Generalized Additive Model

GAM relax the conditions of the GLM. The GAM models takes the form:

$$\boxed{g\left(\mathbb{E}[Y|X_1, ..., X_p]\right) = f_0 + f_1(X_1) + ... + f_p(X_p)}$$

Example: We takes $p = 2$, $g(x) = x$, $f_1(x) = x^2$ and $f_2(x) = x^3$:

$$\mathbb{E}[Y|X_1, X_2] = X_1^2 + X_2^3$$
