Régression logistique.

1- Idée : Reprendre le principe de la régression linéaire pour l'adapter à un problème de classification.

Ds le cas de la régression linéaire, on suppose que la sortie à prédire y (continu) s'écrit $\bar{c}$ une fonct° linéaire / affine des données d'entrée $x = (x_1, .., x_d)^T$:

$$y = f(x) = w_0 + w^T x .$$

avec $w = (w_1, .., w_d)^T$.

Ds le cas de la classification, on ne peut pas faire la m̄ hypothèse puisque y correspond à des classes et donc prend des valeurs discrètes.

1ère idée : Ecrire la probabilité d'obtenir une classe sachant une donnée x $\bar{c}$ une fonction linéaire (vraisemblance) de x :

$$P(y|x) = f(x) = w_0 + w^T x .$$

Mais problème : $P(y|x)$ prend des valeurs entre 0 et 1.

D'où une 2e idée : transformer $f(x)$ pour obtenir des valeurs entre 0 et 1.

$\Rightarrow$ Utilisation de la fonct° sigmoïde.

On s'intéresse au cas binaire : y peut prendre 2 valeurs : 0 et 1.

<u>On note</u> : $P(y=1|x) = \mu(x)$

et $P(y=0|x) = 1-\mu(x)$.

Cas général : $P(y|x) = \mu(x)^y \left(1-\mu(x)\right)^{1-y}$.

2 - <u>Hypothèse</u> :

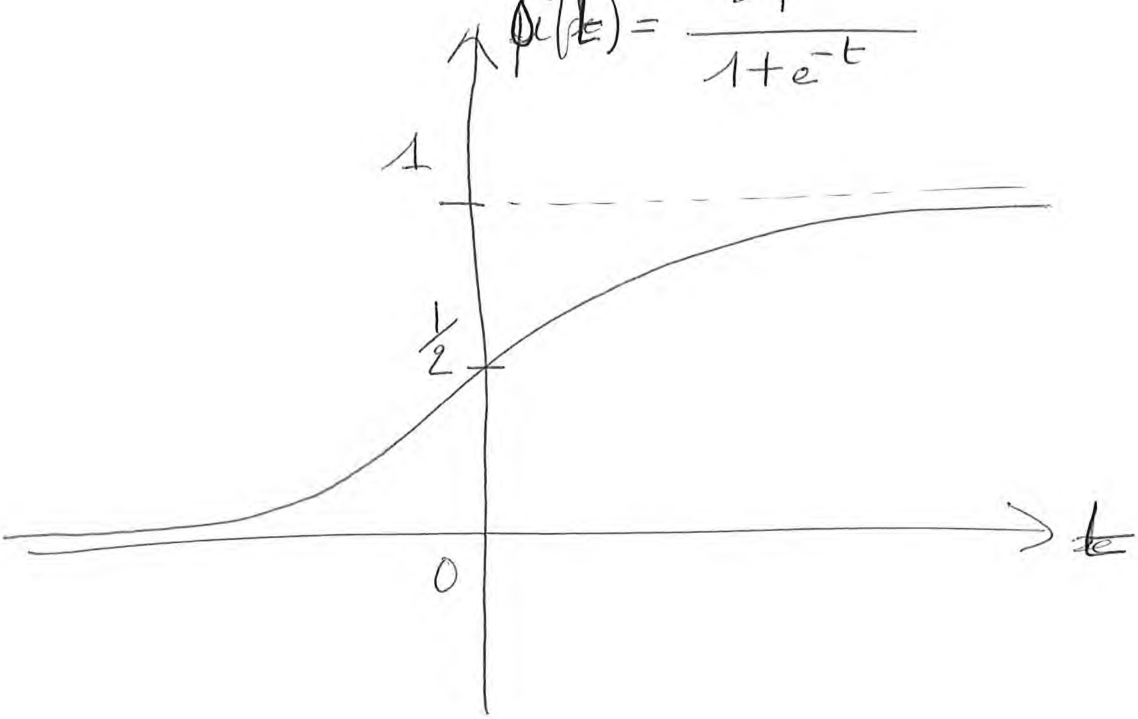
$$\mu(x) = \sigma\left(f(x)\right) \quad \text{avec} \quad f(x) = w_0 + w^T x.$$

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

Avec cette écriture, $\mu(x)$ est compris entre 0 et 1.

$\sigma(\cdot)$ est la fonction sigmoïde

$$\sigma(t) = \frac{1}{1+e^{-t}}$$



Cette hypothèse est équivalente à :

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

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

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

$$\log \frac{1-\mu(x)}{\mu(x)} = -f(x)$$

$$\boxed{\log \frac{\mu(x)}{1-\mu(x)} = f(x) = w_0 + w^T x .}$$

avec $\log \dfrac{\mu(x)}{1-\mu(x)} = \log \dfrac{P(y=1|x)}{P(y=0|x)}$

$\Rightarrow$ fonct° logit.

Finalemt c'est la fonction logit, ie le log-rapport des probabilités qui est supposée s'écrire ē une fonction linéaire de x.

3- Règle de décision :

$$\begin{cases} \text{Si } \mu(x) = P(y=1|x) > \frac{1}{2} \text{ alors } y = 1. \\ \text{Si } \mu(x) < \frac{1}{2} \text{ alors } y = 0. \end{cases}$$

Rem :

$$\mu(x) > \frac{1}{2} \iff \frac{1}{1+e^{-f(x)}} > \frac{1}{2}$$

$$\iff f(x) = w_0 + w^T x > 0.$$

Les 2 classes sont donc partagées par un hyperplan séparateur d'équation $f(x) = 0$. (classification linéaire).

Les paramètres / coefficients $w = (w_1, \dots, w_d)^T$ et w_0 sont à déterminer par l'algorithme.

4 - Objectif de l'algorithme

L'algo cherche (w, w_0) qui maximisent la vraisemblance des n ex. d'apprentissage (principe du max de vraisemblance) : $\prod_{j=1}^{n} P(y^j | x^j)$

$\Leftrightarrow$ maximiser la log-vraisemblance : $\sum_{j=1}^{n} \log P(y^j | x^j)$

Or $P(y^j | x^j) = \mu(x^j)^{y^j} (1 - \mu(x^j))^{1-y^j}$

$\log P(y^j|x^j) = y^j \log \mu(x^j) + (1-y^j) \log(1 - \mu(x^j))$

La log-vraisemblance à maximiser s'écrit :

$$\sum_{j=1}^{n} \left[y^j \log \mu(x^j) + (1-y^j) \log(1-\mu(x^j)) \right]$$

$$\sum_{j=1}^{n} \left[y^j \log \frac{\mu(x^j)}{1-\mu(x^j)} + \log(1-\mu(x^j)) \right]$$

$$\sum_{j=1}^{n} \left[y^j \cdot f(x^j) + \log \Big(\underbrace{1 - \frac{1}{1+e^{-f(x^j)}}}_{\frac{e^{-f(x_j)}}{1+e^{-f(x_j)}} = \frac{1}{1+e^{f(x^j)}}} \Big) \right]$$

$$\sum_{j=1}^{n} y^j \quad \left(w_0 + w^T x^j\right) - \log\left(1 + e^{+\left(w_0 + w^T x^j\right)}\right)\Big]$$

$\Rightarrow$ Pas de solution analytique.

5 - Algorithme.

Maximiser la log-vraisemblance $\Leftrightarrow$
Minimiser la fonction de coût suivante :

$$J(w, w_0) = -\sum_{j=1}^{n} y^j \quad \left(w_0 + w^T x^j\right) - \log\left(1 + e^{+\left(w_0 + w^T x^j\right)}\right)\Big]$$

Une solution consiste à utiliser une technique de descente de gradient pour minimiser $J(w, w_0)$.

Ce sont les paramètres w, w_0 qui sont mis à jour lors de la phase d'apprentissage de l'algorithme.