La fonction logistique en Machine Learning

1 De la régression linéaire à la régression logistique (calculs détaillés)

1.1 Régression linéaire : modèle et estimation

On dispose d'un jeu de données supervisé :

$$\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}, \quad x^{(i)} \in \mathbb{R}^n, \quad y^{(i)} \in \mathbb{R}.$$

Le modèle de régression linéaire est :

$$\hat{y} = w^T x + b.$$

On minimise l'erreur quadratique moyenne :

$$J_{\text{lin}}(w, b) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(w^T x^{(i)} + b - y^{(i)}\right)^2.$$

Pourquoi ce modèle n'est pas adapté à la classification binaire. Si $y \in \{0, 1\}$, alors $\hat{y} = w^T x + b$ peut prendre n'importe quelle valeur réelle. Il n'est donc *pas* interprétable comme une probabilité (qui doit être dans $[0, 1]$). De plus, un modèle linéaire avec perte quadratique n'est pas fondé naturellement sur une loi de Bernoulli.

1.2 Idée centrale : imposer une probabilité via une fonction de liaison

Pour une classification binaire, on cherche à modéliser :

$$p(x) = P(y = 1 \mid x) \quad \text{avec} \quad p(x) \in (0, 1).$$

On conserve un **score linéaire** :

$$z = w^T x + b \in \mathbb{R},$$

puis on le transforme en probabilité par une fonction $g : \mathbb{R} \to (0, 1)$.

1.3 Motivation par les *odds* et le *logit* (dérivation)

On définit les **cotes** (*odds*) :

$$\text{odds}(x) = \frac{P(y = 1 \mid x)}{P(y = 0 \mid x)} = \frac{p(x)}{1 - p(x)}.$$

Les cotes sont dans $(0, +\infty)$, donc on prend le log pour obtenir une quantité dans $\mathbb{R}$:

$$\log(\text{odds}(x)) = \log\left(\frac{p(x)}{1 - p(x)}\right) \in \mathbb{R}.$$

Hypothèse de linéarité (logit linéaire) :

$$\log\left(\frac{p(x)}{1-p(x)}\right) = w^T x + b.$$

On résout pour $p(x)$ (calcul détaillé) :

$$\log\left(\frac{p}{1-p}\right) = z \iff \frac{p}{1-p} = e^z \iff p = e^z(1-p)$$

$$\iff p = e^z - e^z p \iff p + e^z p = e^z \iff p(1+e^z) = e^z$$

$$\iff p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} = \sigma(z).$$

On obtient donc la **fonction logistique** :

$$p(x) = \sigma(w^T x + b), \quad \sigma(z) = \frac{1}{1+e^{-z}}.$$

1.4 Dérivée de la sigmoïde (calcul complet)

Partons de :

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

Dérivons :

$$\sigma'(z) = -1 \cdot (1+e^{-z})^{-2} \cdot \frac{d}{dz}(1+e^{-z}) = -(1+e^{-z})^{-2} \cdot (-e^{-z})$$

$$\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2}.$$

On réécrit en fonction de $\sigma(z)$:

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad \Rightarrow \quad 1-\sigma(z) = \frac{e^{-z}}{1+e^{-z}}.$$

Alors :

$$\sigma(z)(1-\sigma(z)) = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}} = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma'(z).$$

Donc :

$$\boxed{\sigma'(z) = \sigma(z)\big(1-\sigma(z)\big).}$$

2 Construction du coût par maximum de vraisemblance (MLE)

2.1 Modèle probabiliste : Bernoulli conditionnelle

On suppose :

$$y^{(i)} \mid x^{(i)} \sim \text{Bernoulli}(p^{(i)}), \quad p^{(i)} = \sigma(z^{(i)}), \quad z^{(i)} = w^T x^{(i)} + b.$$

La probabilité (pmf) vaut :

$$P\big(y^{(i)} \mid x^{(i)}\big) = \big(p^{(i)}\big)^{y^{(i)}} \big(1-p^{(i)}\big)^{1-y^{(i)}}.$$

2.2 Vraisemblance

Sous indépendance conditionnelle des observations :

$$\mathcal{L}(w,b) = \prod_{i=1}^{m} \big(p^{(i)}\big)^{y^{(i)}} \big(1-p^{(i)}\big)^{1-y^{(i)}}.$$

2.3 Log-vraisemblance (calcul complet)

On prend le log :

$$\ell(w,b) = \log \mathcal{L}(w,b) = \sum_{i=1}^{m} \left[y^{(i)} \log p^{(i)} + (1-y^{(i)}) \log(1-p^{(i)}) \right].$$

On maximise ℓ, équivalent à minimiser la **log-perte** (entropie croisée binaire) :

$$J(w,b) = -\frac{1}{m}\,\ell(w,b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log p^{(i)} + (1-y^{(i)}) \log(1-p^{(i)}) \right].$$

Avec $p^{(i)} = \sigma(w^T x^{(i)} + b)$.

3 Gradients (dérivations pas à pas)

Pour alléger, on fixe une observation et on note :

$$y = y^{(i)}, \quad x = x^{(i)}, \quad z = w^T x + b, \quad p = \sigma(z).$$

La perte (sans moyenne) :

$$\mathcal{L}(w,b) = -\left[y \log p + (1-y) \log(1-p) \right].$$

3.1 Dérivée de $\mathcal{L}$ par rapport à z

On dérive :

$$\frac{d\mathcal{L}}{dz} = -\left[y \frac{1}{p} \frac{dp}{dz} + (1-y) \frac{1}{1-p} \frac{d(1-p)}{dz} \right].$$

Or :

$$\frac{d(1-p)}{dz} = -\frac{dp}{dz}.$$

Donc :

$$\frac{d\mathcal{L}}{dz} = -\left[y \frac{1}{p} \frac{dp}{dz} - (1-y) \frac{1}{1-p} \frac{dp}{dz} \right] = -\frac{dp}{dz} \left[\frac{y}{p} - \frac{1-y}{1-p} \right].$$

Mettons au même dénominateur :

$$\frac{y}{p} - \frac{1-y}{1-p} = \frac{y(1-p) - p(1-y)}{p(1-p)} = \frac{y - yp - p + py}{p(1-p)} = \frac{y-p}{p(1-p)}.$$

Ainsi :

$$\frac{d\mathcal{L}}{dz} = -\frac{dp}{dz} \cdot \frac{y-p}{p(1-p)}.$$

Mais $\frac{dp}{dz} = \sigma'(z) = p(1-p)$, donc :

$$\frac{d\mathcal{L}}{dz} = -(p(1-p)) \cdot \frac{y-p}{p(1-p)} = -(y-p) = p - y.$$

On obtient la relation fondamentale :

$$\boxed{\frac{d\mathcal{L}}{dz} = p - y.}$$

3.2 Gradient par rapport à w

On a $z = w^T x + b$, donc :

$$\frac{\partial z}{\partial w} = x.$$

Par la règle de chaîne :

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{d\mathcal{L}}{dz} \cdot \frac{\partial z}{\partial w} = (p - y)\, x.$$

En réintroduisant la moyenne sur m exemples :

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}^{(i)} \quad \Rightarrow \quad \nabla_w J(w,b) = \frac{1}{m}\sum_{i=1}^{m}(p^{(i)} - y^{(i)})x^{(i)}.$$

Sous forme matricielle, avec $X \in \mathbb{R}^{m\times n}$, $p \in \mathbb{R}^m$, $y \in \mathbb{R}^m$:

$$\boxed{\nabla_w J = \frac{1}{m} X^T (p - y).}$$

3.3 Gradient par rapport à b

On a $\frac{\partial z}{\partial b} = 1$, donc :

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{d\mathcal{L}}{dz} \cdot 1 = p - y.$$

Ainsi :

$$\boxed{\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}(p^{(i)} - y^{(i)}).}$$

4 Pourquoi pas de solution fermée (explication mathématique)

Les conditions d'optimalité imposent :

$$\nabla_w J = \frac{1}{m} X^T (p - y) = 0 \quad \text{et} \quad \frac{\partial J}{\partial b} = 0.$$

Mais $p = \sigma(Xw + b\mathbf{1})$ dépend de w de manière *non linéaire* (exponentielle). Donc, contrairement à la régression linéaire, on n'obtient pas un système linéaire de type

$$X^T X w = X^T y$$

mais un système non linéaire :

$$X^T \sigma(Xw + b\mathbf{1}) = X^T y,$$

qui ne se résout pas en forme fermée en général. Il faut donc utiliser des méthodes numériques itératives.

5 Hessienne, convexité, Newton et IRLS (calculs détaillés)

5.1 Hessienne

On rappelle :

$$\nabla_w J = \frac{1}{m} X^T (p - y), \quad p = \sigma(z), \quad z = Xw + b\mathbf{1}.$$

On calcule la dérivée de p :

$$\frac{\partial p_i}{\partial z_i} = p_i(1 - p_i).$$

On définit la matrice diagonale :

$$R = \operatorname{diag}\left(p_1(1-p_1), \ldots, p_m(1-p_m)\right).$$

Alors :

$$\frac{\partial p}{\partial w} = RX$$

(dimension $m \times n$). Donc :

$$H = \nabla_w^2 J = \frac{1}{m} X^T \left(\frac{\partial p}{\partial w}\right) = \frac{1}{m} X^T R X.$$

Ainsi :

$$\boxed{H = \frac{1}{m} X^T R X.}$$

5.2 Convexité

Comme $0 < p_i < 1$, on a $p_i(1-p_i) > 0$, donc R est définie positive (diagonale à termes positifs). Par conséquent, pour tout vecteur $v \in \mathbb{R}^n$:

$$v^T H v = \frac{1}{m} v^T X^T R X v = \frac{1}{m} (Xv)^T R (Xv) \geq 0.$$

Donc H est semi-définie positive, et J est convexe (souvent strictement convexe si X a rang plein).

5.3 Descente de gradient

Les mises à jour sont :

$$w^{(t+1)} = w^{(t)} - \alpha \nabla_w J(w^{(t)}, b^{(t)}), \quad b^{(t+1)} = b^{(t)} - \alpha \frac{\partial J}{\partial b}(w^{(t)}, b^{(t)}).$$

5.4 Newton-Raphson / IRLS

La méthode de Newton utilise :

$$w^{(t+1)} = w^{(t)} - H^{-1} \nabla_w J.$$

Avec :

$$\nabla_w J = \frac{1}{m} X^T (p - y), \quad H = \frac{1}{m} X^T R X.$$

En négligeant le facteur $\frac{1}{m}$ qui se simplifie :

$$w^{(t+1)} = w^{(t)} - (X^T R X)^{-1} X^T (p - y).$$

Cette méthode est équivalente à une résolution de moindres carrés pondérés itératifs (IRLS).

6 Résumé de la transition Linéaire → Logistique

— Régression linéaire : $\hat{y} = w^T x + b$ (sortie dans $\mathbb{R}$).
— Classification binaire : besoin d'une probabilité $p(x) \in (0, 1)$.
— On suppose un logit linéaire :

$$\log\left(\frac{p}{1-p}\right) = w^T x + b \Rightarrow p = \sigma(w^T x + b).$$

— On apprend (w, b) par MLE :

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log p^{(i)} + (1-y^{(i)})\log(1-p^{(i)})\right].$$

— Pas de solution fermée $\Rightarrow$ méthodes numériques (GD, Newton/IRLS).