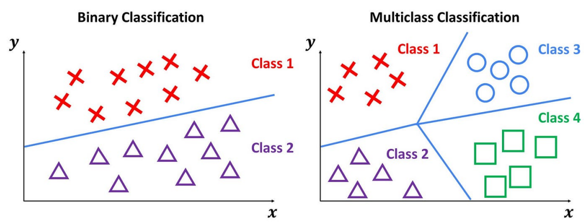
Regression Logistique

Définition et Utilité

- **Objectif :** Modèle de **classification supervisée** pour prédire une probabilité et prendre des décisions binaires.
- **Applications :** Détection de fraude, diagnostic médical, reconnaissance d'objets.
- **Types de classification :** Binaire (0 ou 1), multi-classes (ex. catégories d'objets), multi-étiquettes.

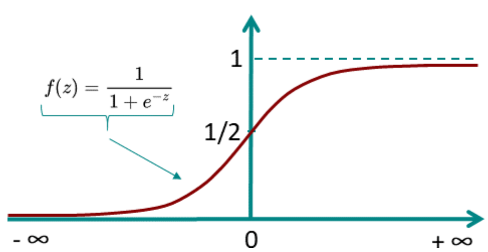


Principe de la Régression Logistique

- **Probabilité** de l'événement $y = 1$ donnée par :

$$p(y = 1|x) = \frac{1}{1 + e^{-(w^T X + b)}}$$

  où w sont les poids et b est le biais.
- **Fonction sigmoïde :** Transforme les valeurs réelles en probabilités entre 0 et 1.



Fonction de Décision

- **Log-odds** (logit) : $\text{score}(x) = \log\left(\frac{p(y=1|x)}{1-p(y=1|x)}\right)$.
- **Décision** : $D(x) = 1$ si $\text{score}(x) \geq 0$, sinon $D(x) = 0$.

Estimation des Paramètres

- Objectif : Trouver les coefficients $\theta = (w, b)$ qui maximisent la **log-vraisemblance**.
- **Vraisemblance** :

$$P(\mathcal{D} \mid \theta) = \prod_{i=1}^{N} P(y_i \mid x_i, \theta) = \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i},$$

  où p_i est la probabilité prédite pour l'observation i, et $\mathcal{D}$ représente les données que l'ont suppose IID.
- **Log-vraisemblance** :

$$L(\theta) = \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

  où p_i est la probabilité prédite pour l'observation i.

Hugo Tondenier

- **Fonction de coût** (entropie croisée) à minimiser :

$$J(\theta) = -\sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Optimisation des Paramètres

- **Descente de gradient** : Algorithme itératif pour minimiser $J(\theta)$.

$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

  où α est le taux d'apprentissage.
- **Méthode de descente** : Dans la pluspart des implémentations, on peut choisir la méthode, par défault c'est Quasi Newton (lbfgs) qui est utilisée
- **Régularisation** : Ajout d'un terme de pénalité (L1 ou L2) pour éviter le surapprentissage.

Prédiction avec le Modèle

- Avec les paramètres optimisés $\hat{\theta}$, la probabilité pour une nouvelle observation x est :

$$p(y = 1|x) = \frac{e^{w^T x + b}}{1 + e^{w^T x + b}}$$

- **Règle de décision** : Si $p(y = 1|x) > 0.5$, prédire $y = 1$, sinon $y = 0$.

Extension à la Multi-classe

- Pour les problèmes multi-classes ($K > 2$), on utilise des **fonctions de score indépendantes** pour chaque classe.
- **Probabilités a posteriori** pour chaque classe k :

$$p(y = k|x) = \frac{e^{w_k^T x}}{1 + \sum_{j=1}^{K-1} e^{w_j^T x}}$$

- **Règle de décision** : Prédire la classe k ayant la probabilité maximale.

Régularisation

- **Objectif :** Éviter le surapprentissage en ajoutant un terme de pénalité à la fonction de coût.
- **Fonction de coût régularisée :**

$$J_{\text{reg}}(\theta) = J(\theta) + \lambda \Omega(\theta)$$

  où $\lambda > 0$ est un hyperparamètre contrôlant l'importance de la régularisation, et $\Omega(\theta)$ est le terme de régularisation.
- **Types de régularisation :**
  - **L2 (Ridge) :**

$$\Omega(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{d} \theta_j^2$$

    Favorise des coefficients plus petits, mais non nuls. Réduit la variance du modèle.
  - **L1 (Lasso) :**

$$\Omega(\theta) = \|\theta\|_1 = \sum_{j=1}^{d} |\theta_j|$$

    Favorise des solutions parcimonieuses où certains coefficients sont exactement nuls, facilitant la sélection des variables.
- **Choix de λ :**
  - λ contrôle le compromis entre l'ajustement du modèle aux données et la pénalisation de la complexité.
  - La **validation croisée** est utilisée pour trouver la valeur optimale de λ.