Régression Logistique

Mini-Cours

Paul Mangold

Audition pour le poste de Maître de Conférences à l'Université Dauphine-PSL

16 mai 2025

Pourquoi la régression logistique ?

- Objectif : Prédire une variable binaire (0 ou 1), ex : malade ou non, succès ou échec.
- Contrairement à la régression linéaire, la régression logistique :
  - Modélise une probabilité.
  - Contraint la sortie à $[0, 1]$.
- Utilisée massivement en médecine, épidémiologie, sciences sociales.

Formulation du modèle

On modélise la probabilité que $Y = 1$ par :

$$P(Y = 1 \mid X) = \sigma(w^T X) = \frac{1}{1 + e^{-w^T X}}$$

où :

- σ est la fonction sigmoïde,
- w est le vecteur de paramètres,
- X est le vecteur de variables explicatives.

Qu'est-ce qu'un *odds ratio* ?

- Définition des **odds** :

$$\text{odds} = \frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)} = \frac{p}{1 - p}$$

- On a alors :

$$\log\left(\frac{p}{1 - p}\right) = w^T X$$

- L'odds ratio (OR) entre deux groupes X et X' vaut :

$$\text{OR} = \exp(w^T (X - X'))$$

Importance des odds ratios en recherche médicale

- Interprétation simple : OR ¿ 1 $\Rightarrow$ le facteur augmente le risque.
- Exemple :
  - OR $= 3.2$ pour un certain médicament $\Rightarrow$ 3.2 fois plus de chances de succès.
- Donne un outil quantitatif pour identifier des facteurs de risque.

Extension au multi-classe : le softmax

Pour K classes :

$$P(Y = k \mid X) = \frac{e^{w_k^T X}}{\sum_{j=1}^{K} e^{w_j^T X}}$$

- Chaque classe k a un vecteur w_k.
- Cette fonction s'appelle **softmax**.
- Utilisé pour généraliser la régression logistique à la classification multi-classes.

La fonction de perte logistique

Pour un seul échantillon (x_i, y_i) :

$$\ell(w; x_i, y_i) = -\left[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)\right]$$

où $\hat{p}_i = \sigma(w^T x_i)$

- Cette fonction pénalise fortement les erreurs de prédiction.
- Convexe et différentiable : propice à l'optimisation par descente de gradient.

Dérivation via maximum de vraisemblance

On suppose que :

$$P(y_i \mid x_i; w) = \sigma(w^T x_i)^{y_i}(1 - \sigma(w^T x_i))^{1-y_i}$$

La vraisemblance totale est :

$$L(w) = \prod_{i=1}^{n} P(y_i \mid x_i; w)$$

On maximise le log :

$$\log L(w) = \sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$$

Donc la **perte logistique** est l'opposée du log-vraisemblance.

Résumé

- La régression logistique modélise $P(Y = 1 \mid X)$ via une fonction sigmoïde.
- L'odds ratio est une mesure clé pour l'interprétation des effets des variables.
- Le softmax permet l'extension à plusieurs classes.
- La fonction de perte logistique provient du maximum de vraisemblance.

Questions ?

Merci de votre attention !