Cours - Informatique - Régression logistique

Gustave Cortal

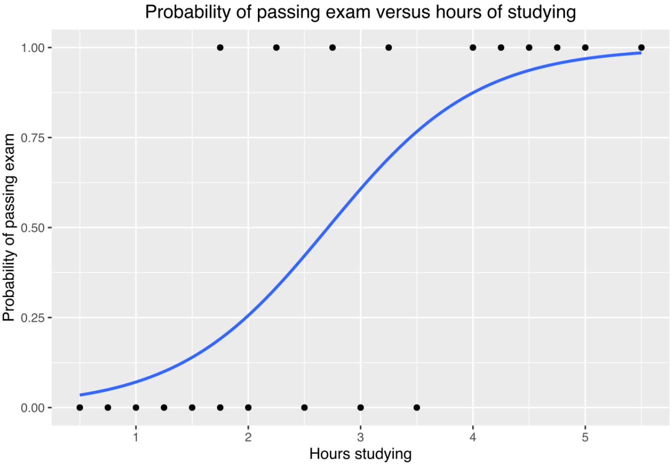
Régression linéaire *vs* logistique

La **régression linéaire** permet d'expliquer une variable **quantitative** (prédiction d'une **quantité**).

La **régression logistique** permet d'expliquer une variable **qualitative** (prédiction d'une **modalité**). Très utilisée pour la prédiction d'une réponse binaire (e.g., *absence*/*présence* d'une pathologie, *achat*/*vente* d'une action, sentiment *positif*/*négatif* d'une phrase).

Comme pour la régression linéaire, la régression logistique prend en entrée des variables explicatives quantitatives ou binaires.

Exemple introductif



Modèle de la régression logistique (1)

L'idée à la base de la régression logistique consiste à modéliser les probabilités *a posteriori* $P(c_k|x)$ par des fonctions de x avec les contraintes $\sum_{k=1}^{g} P(c_k|x) = 1$ et $P(c_k|x) \in [0;1]$ pour tout x.

On va traiter le cas binaire où $g = 2$ classes (c_0 et c_1) avec p variables explicatives.

Pour satisfaire ces contraintes, plusieurs fonctions peuvent être utilisées. On utilise le *modèle logit*, qui consiste à exprimer le logarithme du rapport des probabilités *a posteriori* $P(c_1|x_1, ..., x_p)$ et $P(c_0|x_1, ..., x_p)$ comme une fonction linéaire des x_i.

$$\log\left(\frac{P(c_1|x_1, ..., x_p)}{P(c_0|x_1, ..., x_p)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{1}$$

Modèle de la régression logistique (2)

En observant que $P(c_0|x_1, ..., x_p) = 1 - P(c_1|x_1, ..., x_p)$, on a :

$$P(c_1|x_1, ..., x_p) = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_1 x_1 + ... + \beta_p x_p)} \tag{2}$$

$$P(c_0|x_1, ..., x_p) = \frac{1}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_1 x_1 + ... + \beta_p x_p)} \tag{3}$$

Pour simplifier les notations, on prendra $x = (x_1, ..., x_p)^T$ et $\beta = (\beta_0, ..., \beta_p)^T$ $p_0 = P(c_0|x_1, ..., x_p)$ et $p_1 = P(c_1|x_1, ..., x_p)$

Apprentissage de la régression logistique binaire (1)

On a un ensemble d'apprentissage (x_i, z_i) avec $i = 1, ..., n$. On code la classe z_i par un indicateur binaire :

Si $z_i = c_1, t_i = 1$

Si $z_i = c_0, t_i = 0$

On peut voir t_i comme la réalisation d'une variable $T_i \sim B(p_i)$

La fonction de vraisemblance conditionnelle associée à l'échantillon $T_1, ..., T_n$ est :

$$L(\beta, t_1, ..., t_n) = \prod_{i=1}^{n} P(T_i = t_i) = \prod_{i=1}^{n} p_i^{t_i}(1 - p_i)^{1-t_i} \tag{4}$$

$$\log L(\beta, t_1, ..., t_n) = \sum_{i=1}^{n} (t_i \log p_i + (1 - t_i) \log(1 - p_i)) \tag{5}$$

Apprentissage de la régression logistique binaire (2)

$$\frac{\log L(\beta)}{\beta} = \sum_{i=1}^{n} x_i(t_i - p_i) \tag{6}$$

On veut trouver β qui annule $\frac{\log L(\beta)}{\beta}$, càd :

$$\frac{\log L(\beta)}{\beta} = 0 \tag{7}$$

Il faut résoudre $p + 1$ équations. On ne peut résoudre directement, donc on cherche β qui maximise la vraisemblance des données observées en utilisant un algorithme d'optimisation itératif comme l'*algorithme de Newton-Raphson*.

Mesure d'ajustement : matrice de confusion

On compare les classes prédites avec les classes réelles en donnant un tableau de contingence.

		Predicted	
		0	**1**
Actual	**0**	30	12
	1	8	56

Ici, 30 observations ayant la prédiction "0" sont réellement des "0", alors que 8 observations ayant la prédiction "0" sont en fait des "1".

Interprétation des coefficients

Soit la côte :

$$\frac{P(c_1|x_1,...,x_p)}{P(c_0|x_1,...,x_p)} = exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \quad (8)$$

Si on augmente d'une unité la variable x_i, la côte de la classe c_1 est multipliée par $exp(\beta_i)$. C'est une augmentation si $\beta_i > 0$, c'est une diminution si $\beta_i < 0$, sinon la variable n'a aucun impact si $\beta_i = 0$.

Si $\frac{P(c_1|x_1,...,x_p)}{P(c_0|x_1,...,x_p)} = 2$, alors la côte pour la classe c_1 est de "2 pour 1".

Aller plus loin

- Significativité des coefficients (test de Wald, test du rapport de vraisemblance)
- Régression logistique multinomiale (plus de 2 classes) et ordinale (modalités ordonnées)
- Théorie bayésienne de la décision : analyse discriminante linéaire/quadratique et bayésien naïf
- Méthode des k plus proches voisins