Régression Logistique

Master MLDS - 2017/2018

Université Paris Descartes

Lazhar.labiod@parisdescartes.fr

Introduction

- Le modèle de régression linéaire que nous avons vu jusqu'à présent est un modèle qui s'applique à prédire une variable continue en fonction d'une variable continue

- Il arrive souvent que l'on veuille prédire une variable binaire à partir d'une (ou plusieurs) variable(s) continue(s) (ou nominales), c'est ce que permet de faire la régression logistique $x \in \mathbb{R}, y \in \{0, 1\}$

- Dans le cas où la variable à prédire Y est une variable nominale (catégorique) on parle de classification. Lorsque la variable Y est continue, on parle de régression.

Introduction

- Problème : prédire une variable à deux issues Y = {0, 1} dont l'une est le succès (Y = 1) et l'autre un échec (Y = 0).
- En réutilisant une technique de régression, on peut chercher la probabilité d'obtenir le succès P(Y = 1), il est alors possible de déduire la probabilité de l'échec : P(Y = 0) = 1 − P(Y = 1)
- On peut se munir d'une règle de décision qui pour un seuil θ décide : Y = 1 si P(Y = 1) > θ et 0 sinon

$$Y = \begin{cases} 1 & \text{si} \quad P(Y=1) > \theta \\ 0 & \text{sinon} \end{cases}$$

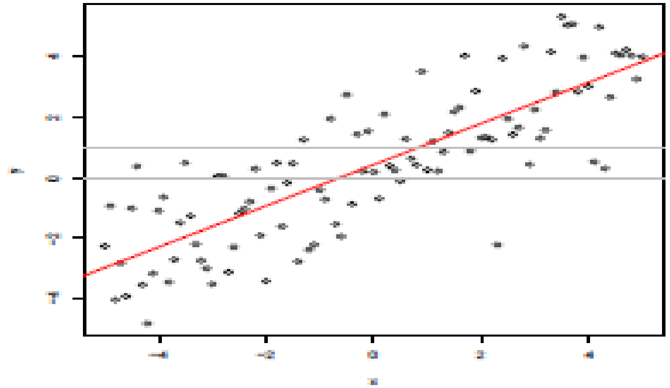
- avec θ = 0.5 en première approximation

Introduction …

- On pourrait envisager réutiliser la régression linéaire pour prédire des valeurs 0 et 1 avec une règle de décision du type :

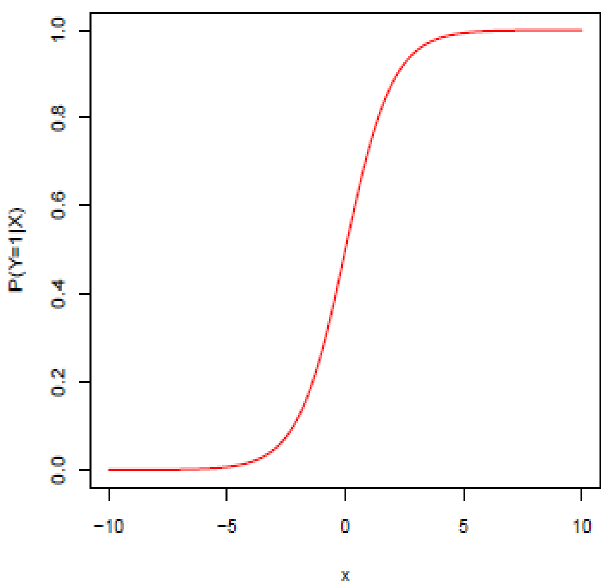
$$Y = \begin{cases} 1 & \text{si} \quad f(x) = \alpha + \beta(x) > 0.5 \\ 0 & \text{sinon} \end{cases}$$

- Le problème est que la régression linéaire produit des valeurs qui sont inévitablement en dehors de l'intervalle $[0, 1] \in \mathbb{R}$ et qui ne s' interprètent pas comme des probabilités
- La régression linéaire va en effet prédire des valeurs continues sur $\mathbb{R}$, or on veut uniquement prédire dans l'intervalle $[0, 1] \in \mathbb{R}$



Fonction logistique

- Les nuages de points dont la variable Y est une variable a valeurs dans [0, 1] ne se résument plus par une droite mais par une fonction qui décrit une courbe en S, la fonction logistique (ou sigmoide)



$$\Pr(Y=1 \mid X=x) = \frac{\exp(\beta_0+\beta x)}{1+\exp(\beta_0+\beta x)},$$

$$\Pr(Y=0 \mid X=x) = \frac{1}{1+\exp(\beta_0+\beta x)}$$

Illustration

- > x <- seq(-10,10,0.1)
- > beta0 <- 1
- > beta <- 1
- > y <- exp(beta0 +beta * x)/(1+ exp(beta0 +beta * x))
- > plot(x,y)
- # *Essayer avec les combinaisons:*
- # beta0 *{-2,0,2}*
- # *beta {-2,-1,1,2}*

Calcul d'une régression logistique

- On a un nuage de points qui se résume par la fonction :

$$\hat{y} = \Pr(Y = 1 \mid X = x) = \frac{\exp(\beta_0 + \beta x)}{1 + \exp(\beta_0 + \beta x)},$$

- Or cette fonction comporte la version linéaire β0 + βx
- On sait comment calculer une régression pour une fonction linéaire
- Idée : projeter la fonction logistique dans un espace linéaire et faire le calcul

Fonction logit

- La fonction logit permet de projeter le problème dans un espace linéaire
- Preuve

$$logit(x) = ln(\frac{x}{1-x})$$

$$
\begin{aligned}
\text{logit}\left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right) &= \ln\left(\frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}\right) \\
&= \ln\left(\frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\frac{1+e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}\right) \\
&= \ln\left(\frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{\frac{1}{1+e^{\alpha+\beta x}}}\right) \\
&= \ln\left(e^{\alpha+\beta x}\right) \\
&= \alpha+\beta x
\end{aligned}
$$

La régression logistique binaire (K=2)

- Les données

Y = variable à expliquer binaire

$X^1, \ldots, X^p$ = variables explicatives numériques ou binaires (indicatrices de modalités)

- Régression logistique simple (p = 1)
- Régression logistique multiple (p > 1)

La régression logistique simple

- Variable dépendante : $Y = 0\ /\ 1$
- Variable indépendante : X
- <u>Objectif :</u> Modéliser

$$\boxed{\pi(x) = \Pr(Y = 1/X = x)}$$

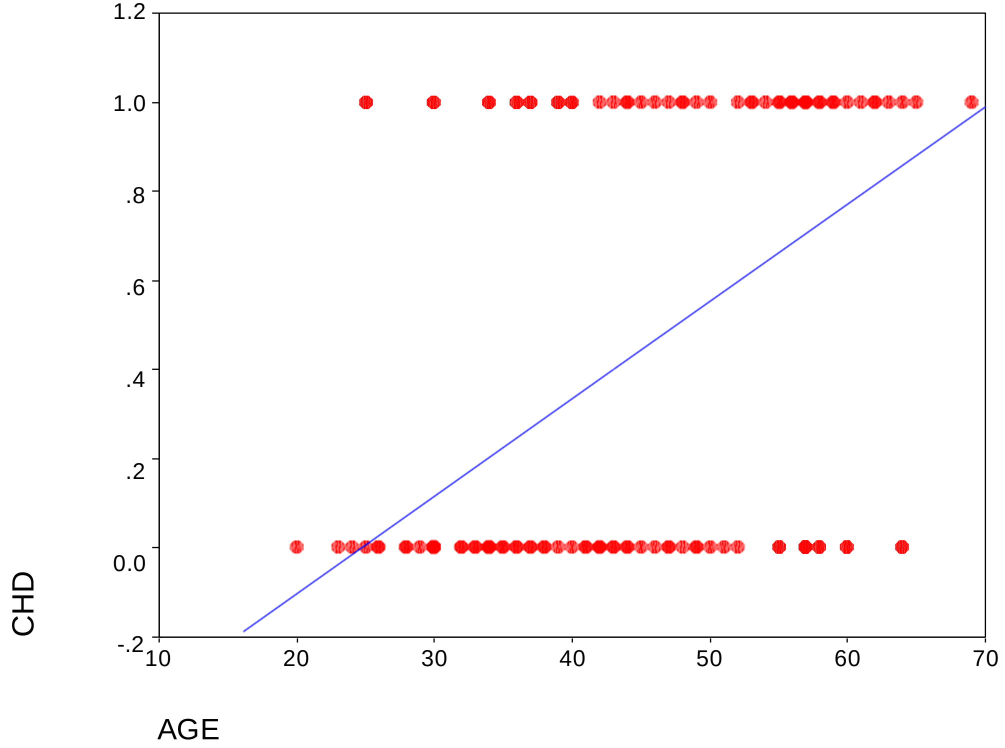
- Le modèle linéaire $\pi(x) = \beta_0 + \beta_1 x$ convient mal lorsque X est continue.
- Le modèle logistique est plus naturel.

Exemple : Age and Coronary Heart Disease Status (CHD)

Les données

ID	AGRP	AGE	CHD
1	1	20	0
2	1	23	0
3	1	24	0
4	1	25	0
5	1	25	1
⋮	⋮	⋮	⋮
97	8	64	0
98	8	64	1
99	8	65	1
100	8	69	1

Plot of CHD by Age



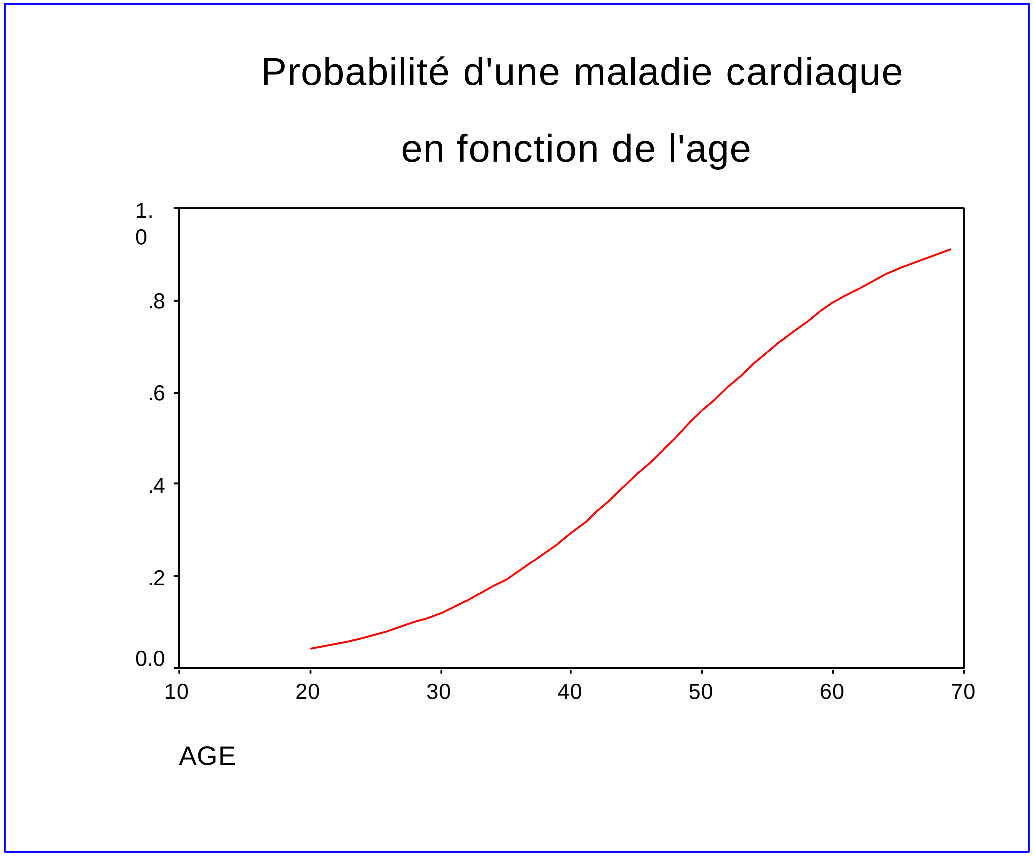
Le modèle logistique

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

ou

$$Log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x$$

Fonction de lien : Logit



Estimation des paramètres du modèle logistique

<u>Les données</u>

X	Y
x_1	y_1
⋮	⋮
x_i	y_i
⋮	⋮
x_n	y_n

y_i = 1 si caractère présent,
0 sinon

<u>Le modèle</u>

$$\pi(x_i) = P(Y = 1 / X = x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Vraisemblance des données

Probabilité d'observer les données

$$[(x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_n,y_n)]$$

$$= \prod_{i=1}^{n} \mathrm{Prob}(Y = y_i \ / \ X = x_i)$$

$$= \prod_{i=1}^{n} \pi(x_i)^{y_i} (1-\pi(x_i))^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)^{1-y_i}$$

$$= \ l(\beta_0,\beta_1)$$

Log-Vraisemblance

$$L(\beta_0,\beta_1) = Log(\ l(\beta_0,\beta_1)) = Log[\prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}]$$

$$= \sum_{i=1}^{n} y_i Log(\frac{\pi(x_i)}{1-\pi(x_i)}) + Log(1-\pi(x_i))$$

$$= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) - Log(1+\exp(\beta_0+\beta_1 x_i))$$

Estimation du maximum de vraisemblance

- On cherche $\hat{\beta}_0$ et $\hat{\beta}_1$ maximisant la Log-vraisemblance $L(\hat{\beta}_0, \hat{\beta}_1)$.

- La matrice

$$V(\hat{\beta}) = \begin{bmatrix} V(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & V(\hat{\beta}_1) \end{bmatrix}$$

est estimée par la matrice $\left[-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} \right]^{-1}_{\beta = \hat{\beta}}$

La régression logistique multi-classe (K>2

La méthode modélise directement les probabilités a posteriori comme des variables à expliquer d'une régression

$$\Pr(Y=k \mid X=x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0} + \beta_l^T x)}, \; k = 1,\ldots, K-1$$

$$\Pr(Y=K \mid X=x) = \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0} + \beta_l^T x)}$$

x une observation (un vecteur p-dimensionnel)

β_k un vecteur p-dimensionnel pour chaque k

Nombre total des paramètres $(K\text{-}1)(p+1)$

Noter que les frontières des classes sont linéaires

Validation du modèle logistique

- Comme pour la régression linéaire, il est important de :

- Vérifier la validité du modèle
- Prouver la qualité de l'ajustement.

Conclusion

- La régression logistique est une méthode de classification trés utilisée, notamment pour K = 2.

- LDA est utile quand n est petit ou les classes sont bien séparées, et l'hypothèse sur la distribution gaussienne des variables est raisonnable. Egalement elle marche bien quand K>2

- Le classifieur bayesien naif est utile quand p est très grand.

Régression logistique versus LDA

- Pour un problèm à deux classes on peut montrer que

$$Log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta x_1^1 + \beta x_2^2 + \ldots + \beta x_p^p$$

- Elle a la même forme que la régrssion logistique
- La différence entre les deux methodes réside dans l'estimation des paramètres
  - La régression logistique utilise la vraissemblance conditionnelle basée sur Pr(Y /X) (une méthode discriminative).
  - LDA utilise la vraissemblance basée sur Pr(X, Y) (une méthode générative).
  - Malgré ces différences , en partique, elles donnent des résultats similaires
- La régression logistique permet également d'ajuster des frontières quadratiques comme la méthode LDA, en injectant explicitement des terms quadratiques dans le modèle