Régression linéaire multiple

Comment la transposer dans un problème de classement ?

Ricco Rakotomalala
Ricco.Rakotomalala@univ-lyon2.fr

Régression vs. Apprentissage Supervisé (Classement)

Régression

Y continue à prédire
X prédictives, quelconques

Classement

Y discrète à prédire
X prédictives, quelconques

On veut construire une fonction de prédiction (explication) telle que

$$Y = f(X, \alpha)$$

Problèmes :
- ☞ il faut choisir une famille de fonction
- ☞ il faut estimer les paramètres α
- ☞ on utilise un échantillon pour optimiser sur la population

Peut-on imaginer un passage de la première approche vers la seconde ?

Critères d'évaluation

Erreur quadratique
Somme des carrés des erreurs

$$S = \sum_{\Omega} [Y - \hat{f}(X, \hat{\alpha})]^2$$

Taux d'erreur
Erreur 0/1 (bon ou mauvais classement)

$$ET = \frac{1}{card(\Omega)} \sum_{\Omega} \Delta[Y, \hat{f}(X, \hat{\alpha})]$$

$$où\, \Delta[.] = \begin{cases} 1\, si\, Y \neq \hat{f}(X, \hat{\alpha}) \\ 0\, si\, Y = \hat{f}(X, \hat{\alpha}) \end{cases}$$

Régression linéaire multiple -- Rappel

- Se restreindre à une famille de fonction de prédiction linéaire
- Et à des exogènes continues (éventuellement des qualitatives recodées)

$$z_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i \ ; \ i = 1, \ldots, n$$

Le terme aléatoire ε cristallise toutes les « insuffisances » du modèle :

- le modèle n 'est qu'une caricature de la réalité, la spécification (linéaire notamment) n 'est pas toujours rigoureusement exacte
- les variables qui ne sont pas prises en compte dans le modèle
- les fluctuations liées à l 'échantillonnage (si on change d 'échantillon, on peut obtenir un résultat différent)

ε quantifie les écarts entre les valeurs réellement observées et les valeurs prédites par le modèle

$(a_0, a_1, \ldots, a_p)$ Sont les paramètres du modèle que 'l'on veut estimer à l'aide des données

Régression vs. Apprentissage supervisé

Mettre en œuvre la régression pour un problème de discrimination binaire

Dans un cas Y binaire (Positifs vs. Négatifs), nous pouvons coder

$$z_i = \begin{cases} 1, \text{si } \mathrm{y_i} = + \\ 0, \text{si } \mathrm{y_i} = - \end{cases}$$

On constate aisément

$$E(Z_i) = P(Y_i = +)$$

Reportée dans l'équation de régression

$$E(Z_i) = P(Y_i = +) = a_0 + a_1 x_{i,1} + \ldots + a_p x_{i,p}$$

On devrait donc pouvoir mettre en place une régression qui permet d'estimer directement la probabilité d'appartenance P(Y=+) ???

hélàs, non…

>> la combinaison linéaire varie entre –∞ et + ∞, ce n'est pas une probabilité
>> les hypothèses de la MCO, notamment l'homoscédasticité et la normalité posent problèmes

Régression vs. Apprentissage supervisé
Séparation linéaire dans l'espace de représentation

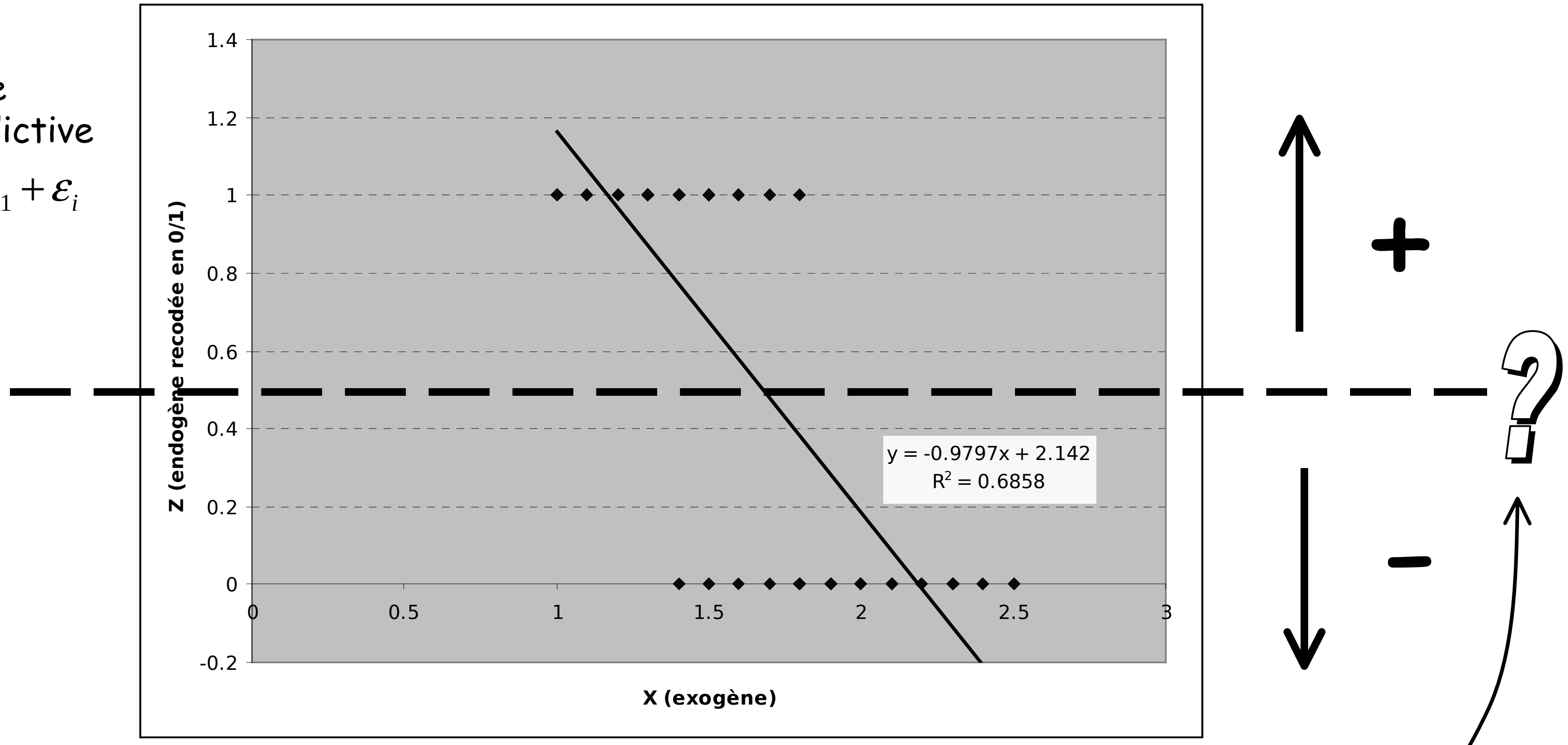
Vue probabiliste

La régression linéaire n'est pas adéquate pour estimer P(Y=+/X), …
Mais on peut exploiter ses résultats pour « séparer » des groupes !!!

Vue géométrique

Ex. Une seule variable prédictive

$$z_i = a_0 + a_1 x_{i,1} + \varepsilon_i$$



Comment définir cette frontière ???

Régression vs. Apprentissage Supervisé

Règle d'affectation (codage 0/1)

Dans un cas Y binaire (Positifs vs. Négatifs), nous pouvons coder

$$z_i = \begin{cases} 1, \text{si } \mathrm{y_i} = + \\ 0, \text{si } \mathrm{y_i} = - \end{cases}$$

On effectue la régression linéaire multiple (DROITEREG dans EXCEL par ex.)

$$z_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i$$

EMCO

$$\hat{z}_i = \hat{a}_0 + \hat{a}_1 x_{i,1} + \hat{a}_2 x_{i,2} + \cdots + \hat{a}_p x_{i,p}$$

Règle d'affectation

$$\hat{y}_i = \begin{cases} +, \text{si } \hat{\mathrm{z}}_\mathrm{i} > \bar{z} \\ -, \text{si } \hat{\mathrm{z}}_\mathrm{i} \leq \bar{z} \end{cases}$$

Moyenne des « z » c.-à-d. $\bar{z} \approx P(Y = +)$

Régression vs. Apprentissage Supervisé
Règle d'affectation (autre codage)

Dans un cas Y binaire (Positifs vs. Négatifs), nous pouvons coder

$$z_i = \begin{cases} \frac{n_-}{n}, \text{si } y_i = + \\ -\frac{n_+}{n}, \text{si } y_i = - \end{cases}$$

On effectue la régression linéaire multiple (DROITEREG dans EXCEL par ex.)

$$z_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i$$

EMCO

$$\hat{z}_i = \hat{a}_0 + \hat{a}_1 x_{i,1} + \hat{a}_2 x_{i,2} + \cdots + \hat{a}_p x_{i,p}$$

Règle d'affectation

$$\hat{y}_i = \begin{cases} +, \text{si } \hat{z}_i > 0 \\ -, \text{si } \hat{z}_i \le 0 \end{cases}$$

On remarque ici

$$\bar{z} = \frac{1}{n}\left(n_+ \times \frac{n_-}{n} + n_- \times (-\frac{n_+}{n}) \right) = 0$$

Régression vs. Analyse Linéaire Discriminante
Analogies et différences dans le cas Y binaire

!

Il est possible de retrouver les coefficients de l'ADL à partir d'une régression linéaire multiple

>> des logiciels utilisent cet artifice en transformant simplement les résultats de la régression

!

Attention, les hypothèses ne sont pas les mêmes, notamment :
• les X sont aléatoires dans la discrimination
• Y est binaire (dist. binomiale) et non pas normale (dist. normale) par l'intermédiaire du terme d'erreur ε

$\Longrightarrow$ Les tests d'évaluation globale (Fisher) et individuel (Student) des coefficients ne peuvent être utilisés qu'à titre indicatif

Comparaison de méthodes

Trois séparateurs linéaires (Rég. Logistique, Analyse Discriminante, Rég. Linéaire)

Régression logistique

$$LOGIT = \ln \frac{P(Y=+/X)}{1-P(Y=+/X)} = a_0 + a_1 x_1 + \cdots + a_p x_p$$

$$\hat{Y} = + \; si \; LOGIT > 0$$

Analyse discriminante linéaire

$$P(Y=+/X) \propto \ln P(Y=+) \times P(X/Y=+) = D(X) = b_0 + b_1 x_1 + \cdots + b_p x_p$$

$$\hat{Y} = + \; si \; D(X) > 0$$

Régression linéaire multiple

$$Z = c_0 + c_1 x_1 + \cdots + c_p x_p \; ; \; Z = \begin{cases} 1, Y = + \\ 0, Y = - \end{cases}$$

$$\hat{Y} = + \; si \; \hat{Z} > \overline{Z}$$

Comparaison de méthodes

Fichier BREAST (Y binaire « cancer », 9 descripteurs continus) – Évaluation en resubstitution

Régression logistique

Attribute	Coef.	Std-dev	Wald	Signif
clump	-0.531	0.132	16.237	0.000
ucellsize	-0.006	0.187	0.001	0.975
ucellshape	-0.333	0.208	2.567	0.109
mgadhesion	-0.240	0.115	4.380	0.036
sepics	-0.069	0.151	0.212	0.645
bnuclei	-0.400	0.089	20.041	0.000
bchromatin	-0.411	0.156	6.918	0.009
normnucl	-0.145	0.102	2.003	0.157
mitoses	-0.551	0.303	3.311	0.069
constant	9.671	-	-	-

	begnin	malignant	Sum
begnin	447	11	458
malignant	11	230	241
Sum	458	241	699

$$\varepsilon = \frac{11+11}{699} = 0.0315$$

Analyse discriminante linéaire

	Classification		Statistical Evaluation			
Attribute	begnin	malignant	Wilks L.	Partial L.	F(1,689)	p-value
clump	0.729	1.616	0.184	0.892	83.767	0.000
ucellsize	-0.316	0.292	0.167	0.983	12.264	0.000
ucellshape	0.066	0.504	0.165	0.990	6.662	0.010
mgadhesion	0.057	0.232	0.164	0.996	2.608	0.107
sepics	0.654	0.870	0.164	0.997	2.290	0.131
bnuclei	0.209	1.427	0.210	0.779	195.186	0.000
bchromatin	0.686	1.245	0.168	0.977	16.553	0.000
normnucl	0.000	0.462	0.169	0.971	20.885	0.000
mitoses	0.201	0.278	0.164	1.000	0.324	0.569
constant	-3.048	-23.296	-			

	begnin	malignant	Sum
begnin	448	10	458
malignant	18	223	241
Sum	466	233	699

$$\varepsilon = \frac{18+10}{699} = 0.0401$$

Régression linéaire multiple (begnin = 1)

Attribute	Coef.	std	t(689)	p-value
clump	-0.033	0.004	-9.152	0.000
ucellsize	-0.023	0.006	-3.502	0.000
ucellshape	-0.016	0.006	-2.581	0.010
mgadhesion	-0.006	0.004	-1.615	0.107
sepics	-0.008	0.005	-1.513	0.131
bnuclei	-0.045	0.003	-13.971	0.000
bchromatin	-0.021	0.005	-4.069	0.000
normnucl	-0.017	0.004	-4.570	0.000
mitoses	-0.003	0.005	-0.569	0.569
Constant	1.253			

	begnin	malignant	Sum
begnin	442	16	458
malignant	4	237	241
Sum	466	233	699

$$\varepsilon = \frac{4+16}{699} = 0.0286$$

Ça n'a pas valeur de preuve...

mais on voit quand même les disparités et les ressemblances

Bibliographique

- *http://fr.wikipedia.org/wiki/Régression_linéaire_multiple*
- *http://fr.wikipedia.org/wiki/Régression_logistique*
- *http://fr.wikipedia.org/wiki/Analyse_discriminante_linéaire*
- *R. Tomassone, M. Danzart, J.J. Daudin, J.P. Masson, « Discrimination et classement », Masson, 1988, pages 36 à 38*
- *G. Saporta, « Probabilités, Analyses de données et Statistique », Technip, 2006, pages 451 et 452.*