

Redressement

Affectation optimale dans le cadre du tirage rétrospectif
Approches analytiques et empiriques

Ricco RAKOTOMALALA

Plan

1. Pourquoi redresser ?
2. Redressement analytique
3. Redressement lors de l'évaluation des performances
4. Redressement empirique – L'utilisation de la courbe ROC et des courbes d'iso-performances
5. Le rapport de vraisemblance, une alternative au critère d'erreur de classement

Démarche de l'apprentissage supervisé

Objectif : Construire une fonction de classement

- Y est la variable à prédire catégorielle
- X sont les prédictives, quelconques

Démarche : Utiliser un échantillon pour l'apprentissage, avec l'espoir d'être performant dans la population → minimisation de l'erreur théorique

N.B.: nous ne tenons pas compte des coûts ici

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

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

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

Schéma bayésien (pour 2 classes $Y = \{+, -\}$:

- Estimer les probabilités $P(Y=+/X)$ et $P(Y=-/X)$
- Prédire $Y=+$ ssi $P(Y=+/X) > P(Y=-/X)$

Ce schéma est repose, entre autres, sur une hypothèse d'échantillonnage aléatoire simple. Est-ce toujours le cas dans la pratique ?



Composition de la population

π_+ Est la probabilité d'être positif ($Y=+$) dans la population

$\pi_- = 1 - \pi_+$ Est la probabilité d'être négatif ($Y=-$) dans la population

Schéma de tirage aléatoire simple (échantillon de taille n)

n_+ (resp. n_-) est le nombre de « positifs » (resp. « négatifs ») dans l'échantillon

$\frac{n_+}{n}$ est bien un estimateur de $P(Y=+)$ dans ce cas

Tirage rétrospectif (échantillon de taille n)

n_+ est fixé a priori, on effectue alors un tirage aléatoire dans le groupe des + (idem pour les « négatifs »)

$\frac{n_+}{n}$ n'est plus du tout un estimateur de $P(Y=+)$, puisque fixé par le statisticien

Pourquoi cette
procédure ?

Souvent utilisée lorsque les « positifs » sont très rares c.-à-d. $P(Y=+)$ est très petit (Ex. les malades, les fraudeurs, les clients de « niche », etc.)

Elle permet d'améliorer la sensibilité des modèles de prédiction

Conséquences du tirage rétrospectif sur le schéma bayésien

Schéma bayésien (pour 2 classes $Y = \{+, -\}$:

Estimer les probabilités $P(Y=+/X)$ et $P(Y=-/X)$

Prédire $Y=+$ ssi (de manière équivalente)

- $P(Y=+/X) > P(Y=-/X)$
- $D = P(Y=+/X) - P(Y=-/X) > 0$
- $R = P(Y=+/X) / P(Y=-/X) > 1$
- $P(Y=+/X) > 0.5$

Or
$$P(Y = + / X) \propto P(Y = +) \times P(X / Y = +)$$



En utilisant n_+/n pour estimer cette probabilité, les calculs usuels sur les logiciels sont faussés

Doit-on :

1. Modifier les calculs de manière à tenir compte explicitement de la vraie valeur de $P(Y=+)$ dans les calculs des probabilités (il y a des logiciels qui peuvent le faire ?) ?
2. Peut-on utiliser les calculs standards des logiciels, puis les corriger par la suite en introduisant l'information sur $P(Y=+)$ (par ex. en modifiant le seuil d'affectation, etc.) ?



Redressement analytique pour l'analyse discriminante

$$\begin{aligned} \text{Score} &= \ln[\pi_+ \times P(X / Y = +)] - \ln[\pi_- \times P(X / Y = -)] \\ &= \ln \frac{\pi_+}{\pi_-} + \ln \frac{P(X / Y = +)}{P(X / Y = -)} \\ &= \ln \frac{\pi_+}{\pi_-} - \ln \frac{n_+}{n_-} + \ln \frac{n_+}{n_-} + \ln \frac{P(X / Y = +)}{P(X / Y = -)} \\ &= \ln \frac{\pi_+}{\pi_-} - \ln \frac{n_+}{n_-} + D \end{aligned}$$

D est fourni en standard par les logiciels, calculé sur la base des effectifs recensés dans le fichier de données (où n_+/n est biaisé)

La règle d'affectation optimale : ($Y = +$) ssi ($\text{Score} > 0$)

est donc équivalente à $(Y = +)$ ssi $\left(D > \ln \frac{n_+}{n_-} - \ln \frac{\pi_+}{\pi_-} \right)$

Commentaires
Importants

1. Nous pouvons exploiter les résultats standards des logiciels, seul le seuil d'affectation est modifié.
2. Cela équivaut à une modification de la constante dans la fonction discriminante
3. Les coefficients associés aux variables (et les tests associés, ex. significativité, etc.) ne sont pas modifiés

Exemple

Prédiction de l'occurrence du diabète (Y) à partir de (X1, X2, X3) les caractéristiques des patients (ex.IMC, etc.).

Données apprentissage : 78 (Y=+) et 78 (Y=-) = 156 obs.

Données test « équilibré » : 50 (Y=+) et 50 (Y=-) (sera utilisé aussi comme *tuning set*)

Données test « représentatif » : 140 (Y=+) et 260 (Y=-)

« Vrai » P(Y=+) = 35 %

Fonction « SCORE » calculée sur les 156 observations en apprentissage

Attribute	Discriminant Function	Classification functions		Statistical Evaluation			
		positive	negative	Wilks L.	Partial L.	F(1,152)	p-value
plasma	0.039233	0.201779	0.162546	0.883321	0.812793	35.00938	0
bodymass	0.073227	0.569753	0.496526	0.758329	0.946763	8.54705	0.00399
pedigree	1.543749	7.400843	5.857094	0.747111	0.960979	6.1721	0.014063
constant	-8.054373	-26.621604	-18.567231			-	

Évaluation sur le fichier test « représentatif » (400 observations, dont 140 Y = +)

Seuil d'affectation « standard » = 0

	positive	negative
positive	95	45
negative	48	212
Taux succès	0.7675	

Seuil d'affectation « corrigé » = 0.62

$$\left(\ln \frac{78}{78} - \ln \frac{0.35}{1-0.35} \approx 0.62 \right)$$

	positive	negative
positive	83	57
negative	25	235
Taux succès	0.795	

Attention, commentaires sur les taux de succès et d'erreur ; et les autres indicateurs (sensibilité, précision, etc....)

Redressement appliqué à l'évaluation des modèles

Le redressement s'applique dans la définition des modèles.

Il s'applique également dans le calcul des ratios
d'évaluation des performances lorsque le fichier test n'est
pas « représentatif ».

Petit rappel sur la Matrice de confusion

Principe : confronter la vraie valeur avec la prédiction

Vrais positifs VP = a

Faux positifs FP = c

$n_+/n = (a+b)/n$

		Prédite		Total
		+	-	
Observée	+	a	b	a+b
	-	c	d	c+d
	Total	a+c	b+d	n

Quelques ratios :

- Taux d'erreur = $(c+b)/n$
- Taux de succès = $(a+d)/n$
- Sensibilité = Rappel = Taux de VP = TPR = $a/(a+b)$
- Précision = $a/(a+c)$
- Taux de FP = $c/(c+d) = \text{FPR}$
- Spécificité = $d/(c+d) = 1 - \text{Taux de FP}$

(1) TPR et FPR ne sont pas affectés par la proportion n_+/n

(2) Les autres indicateurs, y compris le taux d'erreur et le taux de succès, sont dépendants de cette proportion.

(3) Comme n_+/n n'est pas le reflet de $P(Y=+)$ dans la population, le taux d'erreur et le taux de succès calculés ici n'indiquent pas non plus la performance du modèle dans la population.

Redressement lors de l'évaluation

Calcul du « vrai » taux de succès à partir d'une matrice de confusion

$$\theta = \pi_+ \times tpr + (1 - \pi_+) \times (1 - fpr)$$

Remarques : 1. Vérifier que si tirage aléatoire, on retrouve les formules usuelles du taux de succès
2. $P(Y=+)$ peut être obtenu de différentes manières (ex. expertise, autres études, etc.)

Calcul du « vrai » taux de succès à partir d'une matrice de confusion –
Fichier test « représentatif »

Matrice de confusion		
	positive	negative
positive	83	57
negative	25	235
Taux succès	0.795	

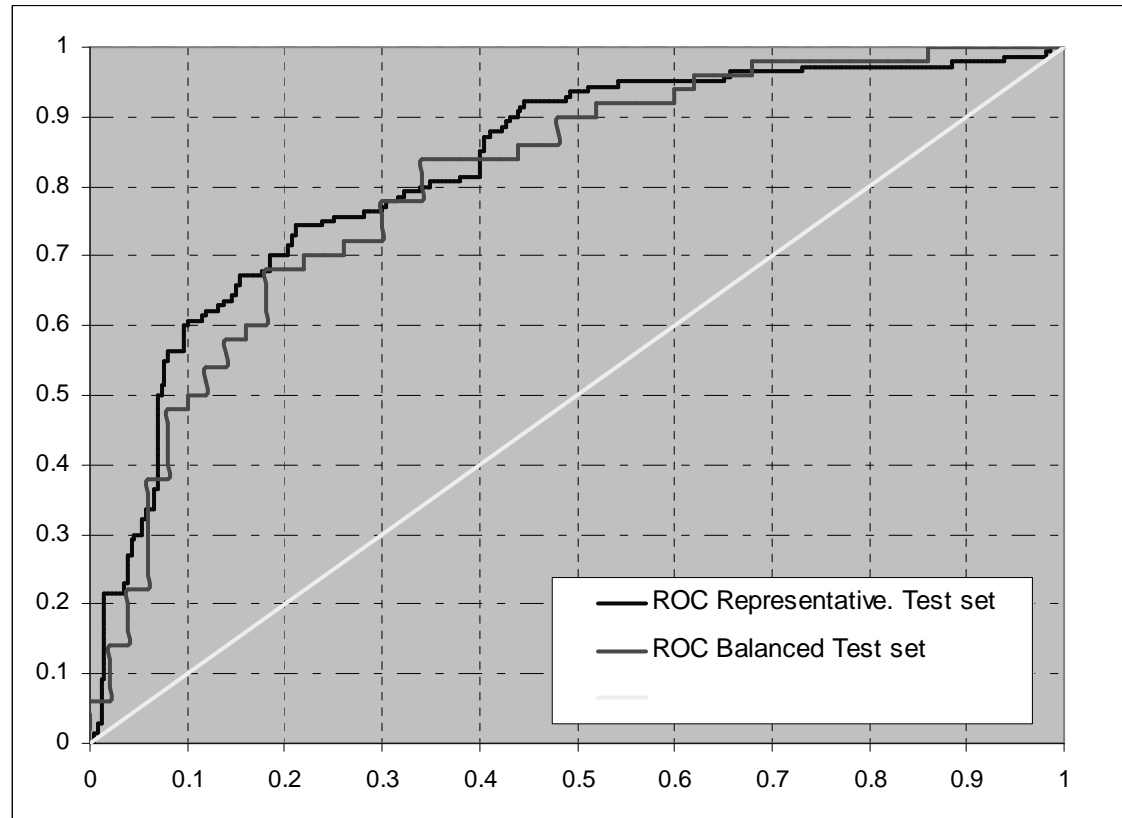
$$0.35 \times \frac{83}{140} + (1 - 0.35) \times \left(1 - \frac{25}{260}\right) = 0.795$$

Calcul du « vrai » taux de succès à partir d'une matrice de confusion –
Fichier test « non-représentatif »
(de même distribution que le fichier apprentissage par ex.)
ex. 50 (Y=+) et 50 (Y=-) = 100 obs.

Matrice de confusion		
	positive	negative
positive	30	20
negative	5	45
Taux apparent	0.75	

$$0.35 \times \frac{30}{50} + (1 - 0.35) \times \left(1 - \frac{5}{50}\right) = 0.795$$

Remarque : la courbe ROC est insensible à la distribution des classes dans le fichier test



Que ce soit sur l'échantillon test représentatif ou l'échantillon test équilibré (de même répartition que l'échantillon d'apprentissage ici), la courbe ROC reste (approximativement – aux fluctuations d'échantillonnage près) la même. Il en est de même pour le critère AUC.

Redressement

Approches empiriques

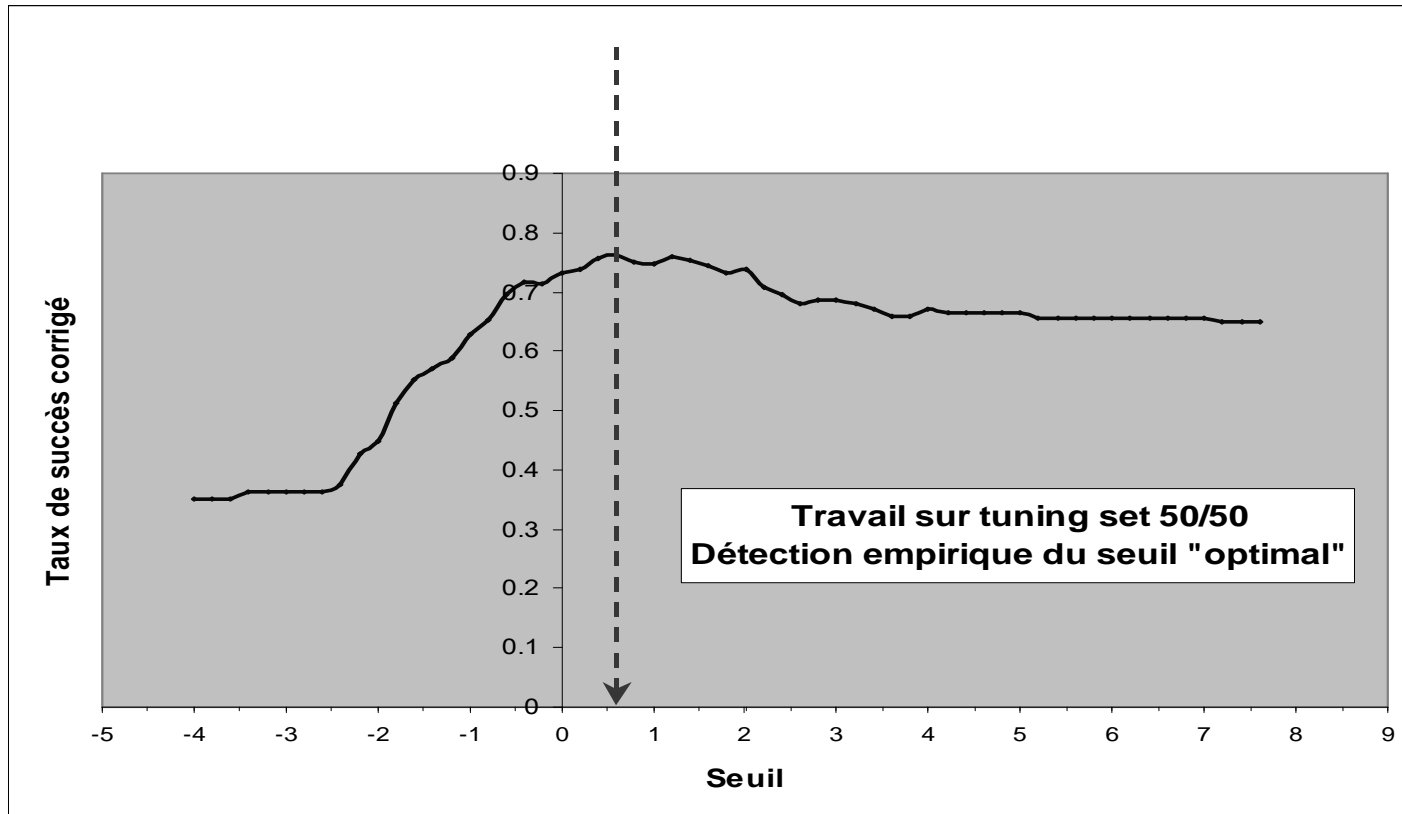
Parfois, selon les méthodes, le seuil « corrigé » ne peut pas être calculé analytiquement. On peut alors s'appuyer sur des démarches empiriques.

Parfois également, nous ne voulons pas optimiser le taux d'erreur mais d'autres indicateurs, plus pertinents dans le contexte de notre étude.

Détection empirique du seuil d'affectation d'optimal (1/2)

Utilisation de la courbe (Seuil vs. Taux de succès)

Démarche : Faire varier le seuil et surveiller le taux de succès corrigé
Attention : Risque de sur-apprentissage si on utilise le fichier d'apprentissage



Il faut alors subdiviser le fichier en 3 parties (*ce n'est pas toujours possible, hélas...*) :

- learning set, pour construire le modèle (ex. 156 obs.)
- tuning set, pour détecter le seuil « optimal ». Il participe à l'apprentissage au final (ex. 100 obs.)
- test set, pour évaluer « honnêtement » les performances (ex. 400 obs.)

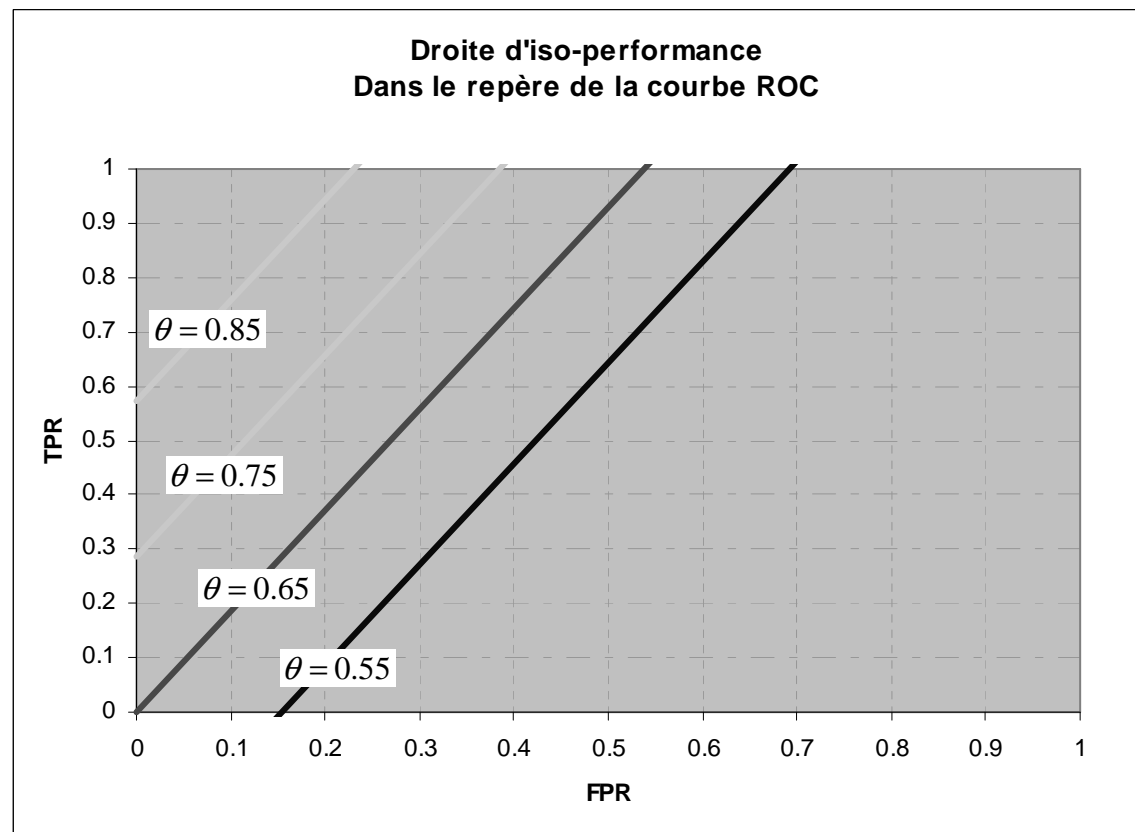
Détection empirique du seuil optimal (2/2)

Utilisation de la courbe ROC et des droites d'iso-performances (1)

$$\theta = \pi_+ \times tpr + (1 - \pi_+) \times (1 - fpr)$$

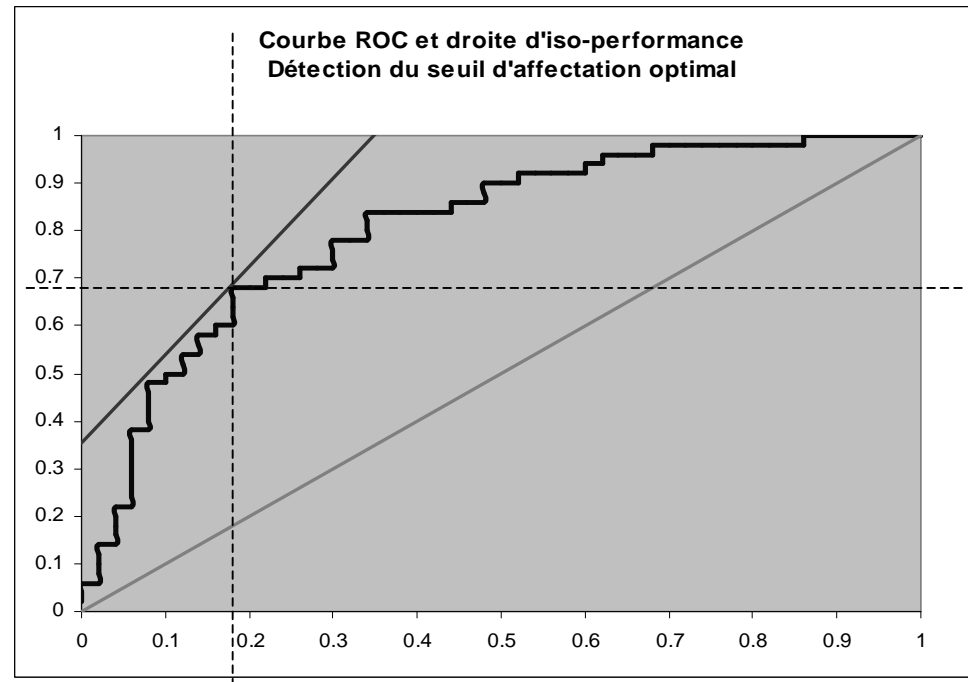
Cette équation, pour π_+ [P(Y=+)] donné, définit une droite dans le repère de la courbe ROC (FPR, TPR).

Tout couple (TPR, FPR) situé sur la même droite, aboutit au même taux de succès !



Détection empirique du seuil optimal (2/2)

Utilisation de la courbe ROC et des droites d'iso-performances (2)



Stratégie de détection, on travaille sur le fichier TUNING SET

1. Construire la courbe ROC
2. Tracer la droite d'iso-performance
3. Faire varier le taux de succès de manière à ce que la droite affleure la courbe ROC (ici : THETA=0.77)
4. Détecter le couple de valeurs (FPR, TPR) correspondant au point de tangence entre la droite d'iso-perf. et la courbe ROC
5. Rapporter ces valeurs dans le tableau de construction de la courbe ROC pour déterminer le score associé : ce sera le seuil d'affectation à utiliser.
6. Le taux THETA optimal trouvé est un estimateur du taux de succès

Tableau de construction de la courbe ROC – Détection du seuil

diabete	score	FPR	TPR
positive	7.093	0	0.02
positive	5.005	0	0.04
...
positive	0.643	0.18	0.64
positive	0.633	0.18	0.66
positive	0.557	0.18	0.68
...
negative	-2.390	0.96	1
negative	-2.403	0.98	1
negative	-3.518	1	1

Autres critères

Se démarquer du taux de succès (d'erreur)

Le taux de succès n'est pas toujours approprié.

Ex. lorsque les classes sont très déséquilibrées -- c.-à-d. les positifs sont très rares -- la meilleure stratégie au sens du taux de succès est de toujours prédire « négatif » ☹

Plus souvent, on s'intéresse à la « propension à être positif » (le score est un outil privilégié pour cela), il nous faut des critères d'évaluation qui retraduisent cette idée (ex. la courbe ROC et AUC permettent déjà de le faire, *what else ?*)

Le rapport de vraisemblance (1/2)

Le rapport de vraisemblance « positif » est défini par :

$$\begin{aligned} LR_+ &= \frac{TPR}{FPR} \\ &= \frac{P(\hat{Y} = + / Y = +)}{P(\hat{Y} = + / Y = -)} \end{aligned}$$

Pré-multiplié par le rapport (la cote) des probabilités a priori, on obtient :

$$\begin{aligned} &\frac{P(Y = +)}{P(Y = -)} \times \frac{P(\hat{Y} = + / Y = +)}{P(\hat{Y} = + / Y = -)} \\ &= \frac{P(\hat{Y} = +, Y = +)}{P(\hat{Y} = +, Y = -)} \\ &= \frac{P(\hat{Y} = +) \times P(Y = + / \hat{Y} = +)}{P(\hat{Y} = +) \times P(Y = - / \hat{Y} = +)} \\ &= \frac{P(Y = + / \hat{Y} = +)}{P(Y = - / \hat{Y} = +)} \end{aligned}$$

Résultat très intéressant, lorsque notre prédit qu'un objet est positif avec notre modèle, le seuil que l'on s'est choisi et la proportion véritable de positifs dans la population, il y a 3.32 fois plus de chances qu'il soit réellement un positif !!! C'est plutôt bon signe.

(1) Il est indépendant de la proportion des ($Y=+$) dans l'échantillon, donc facilement transposable d'une étude à l'autre, que ce soit le mode de constitution des échantillon (aléatoire ou rétrospectif)

(2) Interprétation : Il indique le rapport entre la probabilité d'être classé positif selon que l'on soit réellement positif ou négatif. Plus la valeur est élevée, mieux c'est, cela indique que les positifs ont une plus forte probabilité d'être classés comme tel que les négatifs.

(3) Utilisation : Pré-multiplié par le rapport des probas a priori, il permet d'obtenir le rapport (la cote) des probas a posteriori c.-à-d. la propension à être réellement un positif lorsque la prédiction du modèle est « positif ». Plus cette cote est élevée (>1), plus la prédiction sera fiable.

Matrice de confusion		
	positive	negative
positive	83	57
negative	25	235
Taux succès	0.795	

$$\rightarrow \frac{83/140}{25/260} = 6.16$$

$$\frac{P(Y = + / \hat{Y} = +)}{P(Y = - / \hat{Y} = +)} = \frac{0.35}{0.65} \times \frac{83/140}{25/260} = \frac{0.35}{0.65} \times 6.16 = 3.32$$

Le rapport de vraisemblance (2/2)

Seuil optimal au sens de la « cote » de $P(Y=+ / X)$

Matrice de confusion		
	positive	negative
positive	38	102
negative	11	249
LR+	6.42	
TPR	0.27	
FPR	0.04	
Précision	0.78	
Taux de succès	0.72	

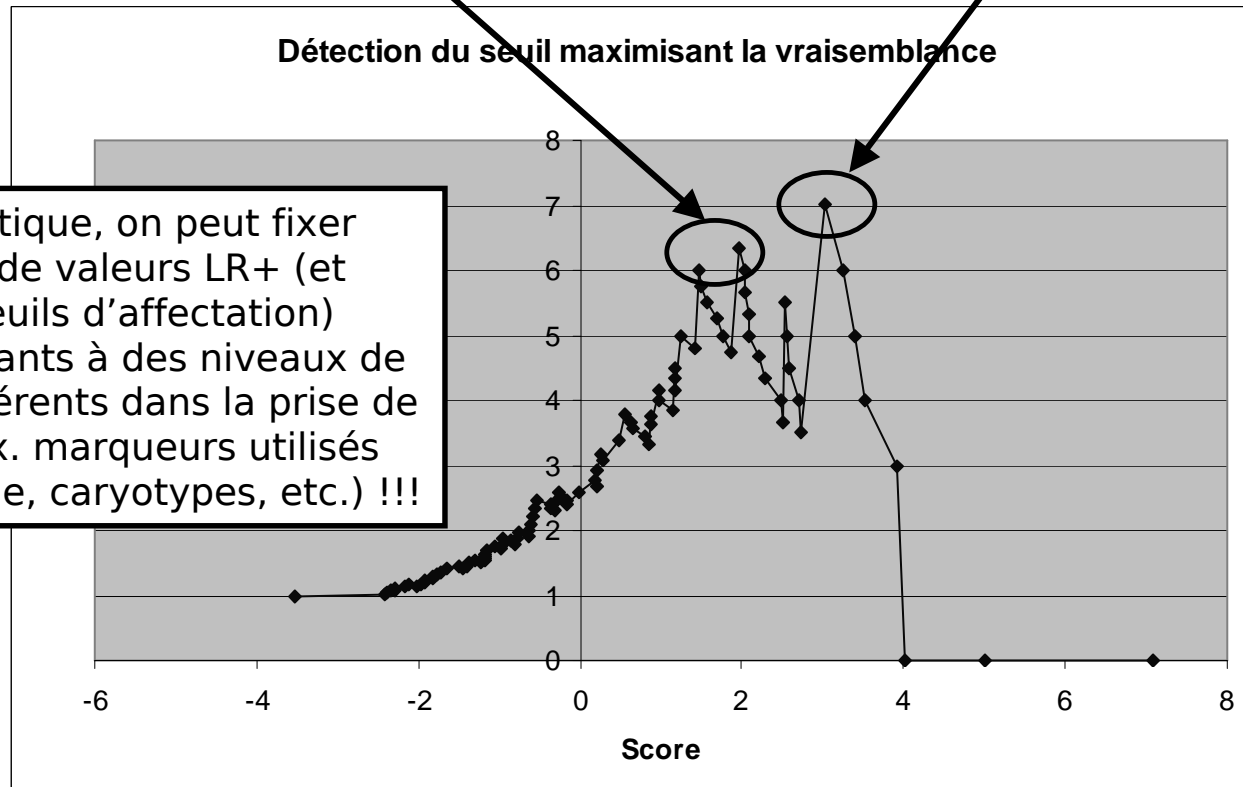
Matrice de confusion		
	positive	negative
positive	13	127
negative	3	257
LR+	8.05	
TPR	0.09	
FPR	0.01	
Précision	0.81	
Taux de succès	0.68	

seuil ≈ 1.8

seuil ≈ 3

Détection du seuil maximisant la vraisemblance

Dans la pratique, on peut fixer des plages de valeurs LR+ (et donc des seuils d'affectation) correspondants à des niveaux de risques différents dans la prise de décision (ex. marqueurs utilisés en médecine, caryotypes, etc.) !!!



Courbe construite sur le « tuning set » toujours.
On perçoit deux solutions ici, que nous évaluons/confrontons sur le « test set ».

Références

Formule de correction analytique pour la régression logistique
M. Bardos, « Analyse Discriminante – Application au risque et au scoring financier », Dunod, 2001.

Quelques considérations autour de la courbe ROC
T. Fawcett, « ROC Graphs : Notes and Practical Considerations for Researchers », TR-HP Laboratories, 2004.

Comparaisons ROC, LIFT etc. -- Affectation optimale
M. Vuk, T. Curk, « ROC Curve, LIFT Chart and Calibration Plot », Metodoloski Zvezki, vol.3, n°1, pp.89-108, 2006.

Conclusion

Dès que l'on veut transposer les résultats dans la population dans des études réelles, il faut toujours étudier le mode d'élaboration du fichier de données. Il faut s'assurer notamment, que la distribution empirique des classes est représentative des probabilités a priori.

Le cas échéant, il faut penser à redresser les résultats avant de procéder aux affectations et calculer les indicateurs de performances dépendants de la répartition apparente des classes.

Ou alors, il faut utiliser les outils d'évaluation insensibles à la distribution empirique des classes dans les fichiers utilisés (ex. la courbe ROC)

Mais, de toute manière...

Bien souvent le taux d'erreur ou le taux de succès sont de mauvais indicateurs lorsque les classes sont très déséquilibrées. En effet, lorsque les positifs sont très rares, la « meilleure » décision est de prédire tous les individus négatifs au sens du taux d'erreur → ce n'est pas très satisfaisant...

Dans le contexte des classes très déséquilibrées, la capacité à placer les positifs devant les négatifs est le critère déterminant (cf. courbe LIFT ou encore une fois la courbe ROC et le rapport de vraisemblance).