Normalisation des scores

Proposer une estimation fiable de P(Y=+/X)
dans un problème de discrimination

Ricco RAKOTOMALALA

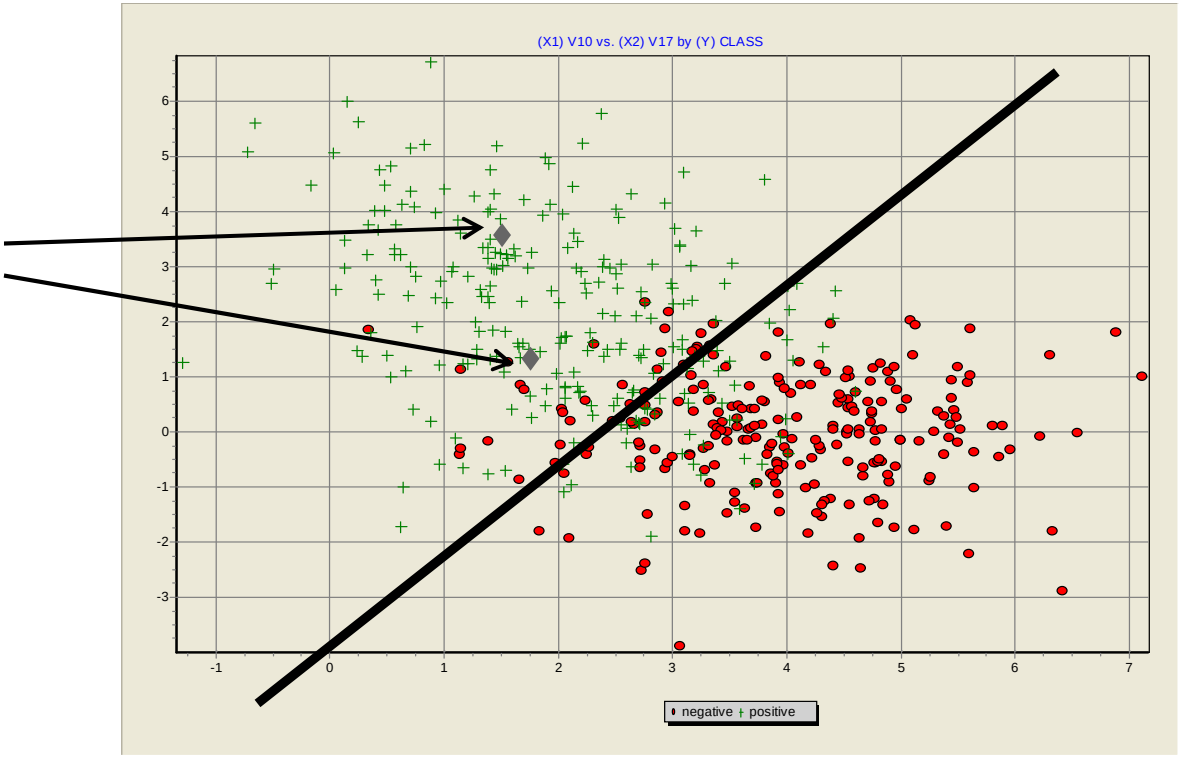
Plan

1. Pourquoi normaliser les scores ?
2. Évaluation de la qualité d'un score
3. Méthode de Platt
4. Méthode de la Régression Isotonique

Classement et crédibilité du classement

On se place dans le cadre binaire dans ce cours

Les deux points sont classés « positifs », mais l'un est plus positif que l'autre !



Un indicateur de crédibilité de l'affectation est souhaitable : on parle souvent de SCORE

• on peut le prendre ici comme un indicateur du degré d'éloignement par rapport à la frontière, il peut être lu autrement dans d'autres cas.

• Il est proportionnel à P(Y=+/X) ç.-à-d. le score permet de classer les observations de la même manière que P(Y=+/X)

• il n'est pas « calibré » ç.-à-d. il varie a priori entre –oo et +oo et peut prendre des plages de valeurs différentes d'un classifieur à l'autre

Problème du score non-calibré

Le problème des disparités entre les plages de valeurs

(1) Le SCORE brut est suffisant pour le classement : SCORE > 0 → Y = +
(2) Il est adapté également au scoring puisqu'il s'agit simplement de trier la base de données selon l'appétence

… mais n'est pas approprié dans la plupart des autres contextes

Interprétation des résultats et déploiement

Pour un expert du domaine, que veut dire un SCORE = 28 ?
Il faut des « vraies » probabilités dès que l'on veut appliquer des coûts

Comparaison des classifieurs

Si la méthode A fournit SCORE = 3 et la méthode B = -1, laquelle croire ?

Combinaison des classifieurs

Si on veut faire coopérer plusieurs méthodes, éviter qu'une d'entre elles, avec ses plages de valeurs élevées, « écrase » les autres.

Passage du classement binaire au multi-classes

Pour les méthodes fondamentalement binaires (ex. SVM, Reg. Logistique, etc.), le passage à K classes passe par la combinaison d'une série de classifieurs binaires avec différentes stratégies de combinaison (ex. 1-vs-All, 1-vs.-1, ECOC, etc.) basé sur les scores, il faut avoir des scores comparables.

Problème du score non-calibré

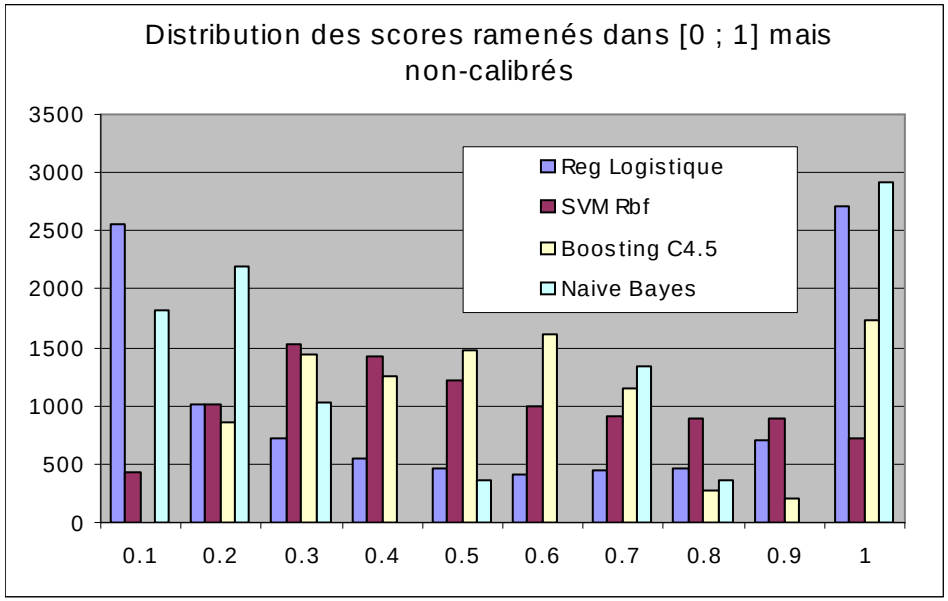
Ramener le score dans l'intervalle [0 ; 1] ne suffit pas

Une technique simple pour ramener les scores dans le même intervalle [0 ; 1]

$$p = \frac{s - s_{\min}}{s_{\max} - s_{\min}}$$

Mais si le problème des plages de valeurs est résolu, le problème de la disparité des distributions des valeurs demeure…

Les comparaisons ne sont pas possibles !



Problème du score non-calibré

Y a-t-il des méthodes naturellement bien calibrées ?

Les méthodes fournissant des scores en adéquation avec P(Y=+/X)

La plupart des méthodes qui proposent des estimations de P(Y=+/X) sans pour autant reposer sur des hypothèses trop contraignantes

- Régression Logistique
- Analyse Discriminante
- Random Forest + Bagging Arbres
- Perceptron (RNA)
- Méthodes à base de noyaux (estimation non-paramétrique des probabilités)

Les méthodes fournissant des scores systématiquement extrêmes (proches de 0 ou 1 lorsqu'ils sont ramenés dans [0 ; 1])

- Bayesien naïf (modèle d'indépendance conditionnelle)

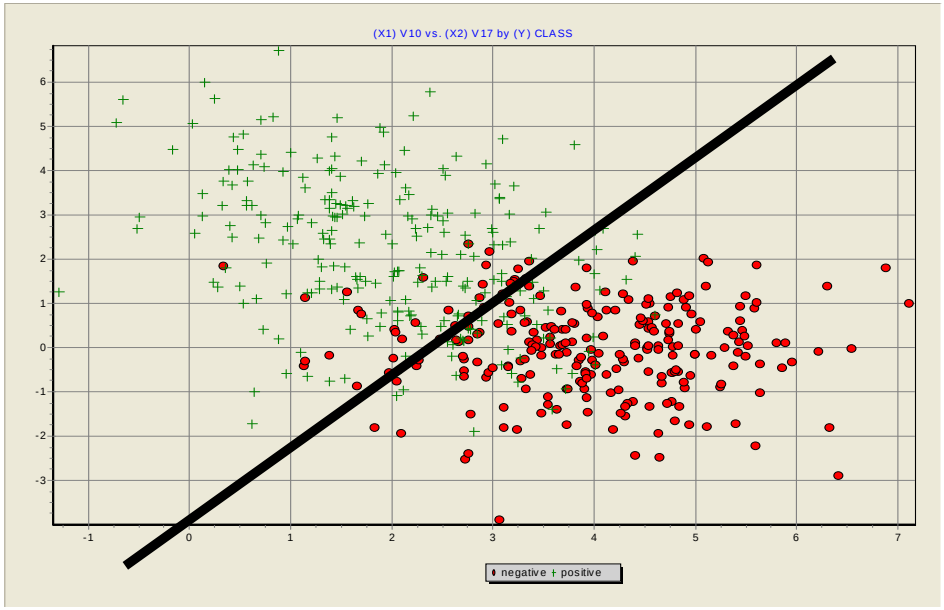
Les méthodes fournissant des scores agglutinés autour de la valeur centrale (0.5 lorsqu'ils sont ramenés dans [0 ; 1]

Les méthodes maximisant la marge

- SVM
- Boosting Arbres

Fichier de travail et démarche de normalisation (« *calibration* », en anglais) des scores

Fichier « WAVE » expurgée de la 3ème classe – Les classes restantes sont équilibrées



Training Set (500 obs.)

Il sert à construire les modèles de prédiction (Régression logistique, SVM Rbf, Boosting C4.5 et Naive Bayes). C'est à partir de ces classifieurs que l'on peut produire les scores non-calibrés.
500 obsv.

Tuning Set (500 obs.)

Il sert à normaliser les scores. Il s'agit donc d'une sorte de deuxième fichier d'apprentissage puisqu'il sert à calculer ou à déterminer les valeurs « optimales » de paramètres. Participant en partie à l'apprentissage, il ne peut être utilisé pour évaluer les performances de la technique ou du dispositif mis en place.

Test Set (10000 obs.)

Il sert uniquement à évaluer le(s) dispositif(s) mis en place.

Évaluation de la qualité d'un score

Un bon score doit varier entre 0 et 1 et être une bonne estimation de P(Y=+/X)

Le diagramme de fiabilité (Reliability Diagram, DeGroot & Fienberger, 1982)

Idée : Vérifier que dans chaque intervalle de scores (ex. 0.0 – 0.2, 0.2 – 0.4, …, 0.8 – 1.0), la proportion des positifs est en adéquation avec la valeur moyenne du score. Si oui, le score est bien calibré, c'est un bon estimateur de P(Y=+/X).

Démarche : Travailler sur le fichier test

- Appliquer le modèle pour obtenir le score de chaque observation
- Trier le fichier selon le score croissant
- Subdiviser les données en quantiles (ex. déciles, quartiles, etc.)
- Dans chaque intervalle, calculer la proportion des positifs et calculer la moyenne des scores
- Si les chiffres concordent dans tous les intervalles → le score est bien calibré

Ex. Régression Logistique (10.000 individus en test)

Proportion de positifs dans chaque intervalle de score

NB CLASS	CLASS			
Plage	negative	positive	Total	P(Y=+)
0.0 - 0.2	3244	309	3553	8.70%
0.2 - 0.4	912	346	1258	27.50%
0.4 - 0.6	473	402	875	45.94%
0.6 - 0.8	278	630	908	69.38%
0.8 - 1.0	137	3269	3406	95.98%
Total	5044	4956	10000	

} X

Moyenne des scores des observations dans chaque intervalle

Y {

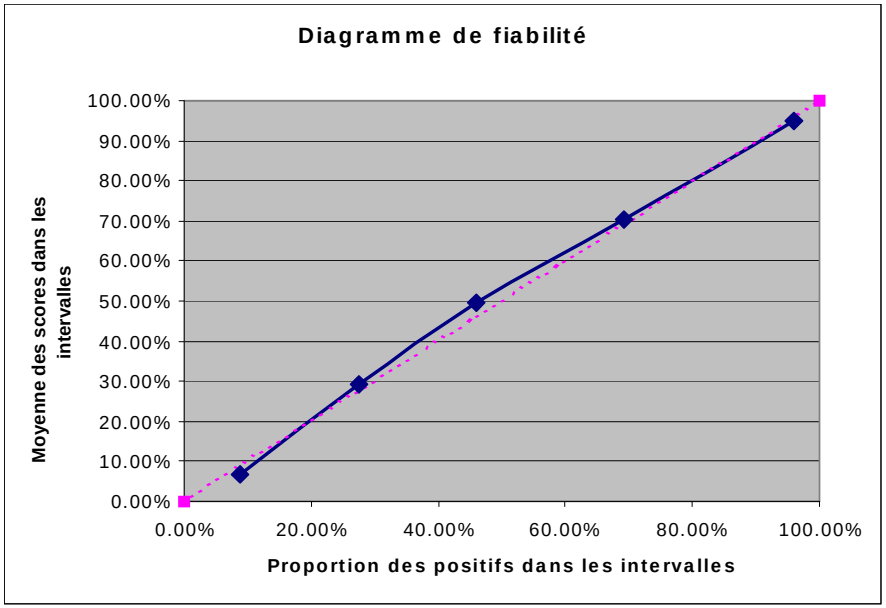
Moyenne SRegLog	
Plage	
0.0 - 0.2	6.91%
0.2 - 0.4	29.14%
0.4 - 0.6	49.53%
0.6 - 0.8	70.33%
0.8 - 1.0	94.94%

Le diagramme de fiabilité est un graphique XY avec en abscisse la proportion de « + » et en ordonnée la moyenne des scores

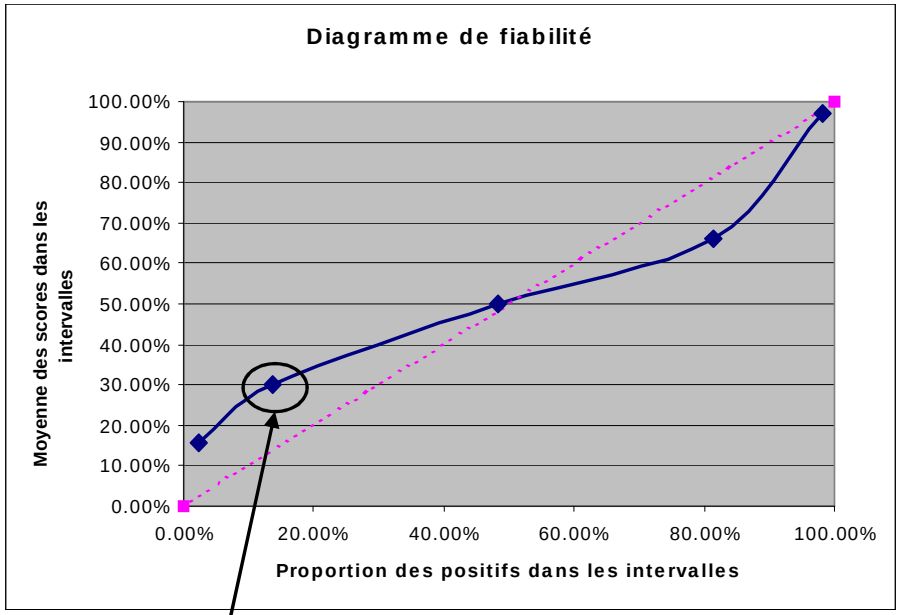
Évaluation de la qualité d'un score

Diagramme de fiabilité -- Distorsion selon les méthodes

Une méthode fournit des scores bien calibrés si elle permet d'élaborer un diagramme proche de la diagonale principale.



Régression Logistique
Score bien calibré



Boosting C4.5
Score mal calibré (ex. intervalle 0.2 – 0.4)

Ex. Dans le deuxième intervalle (0.2 – 0.4), la moyenne des scores est de 30%, et la proportion des positifs dans ce même intervalle est de #16%

Normalisation des scores – Méthode de Platt (1999)

Utiliser la fonction sigmoïde

Idée : Une « bonne » fonction de répartition prend « souvent » une forme sigmoïde.

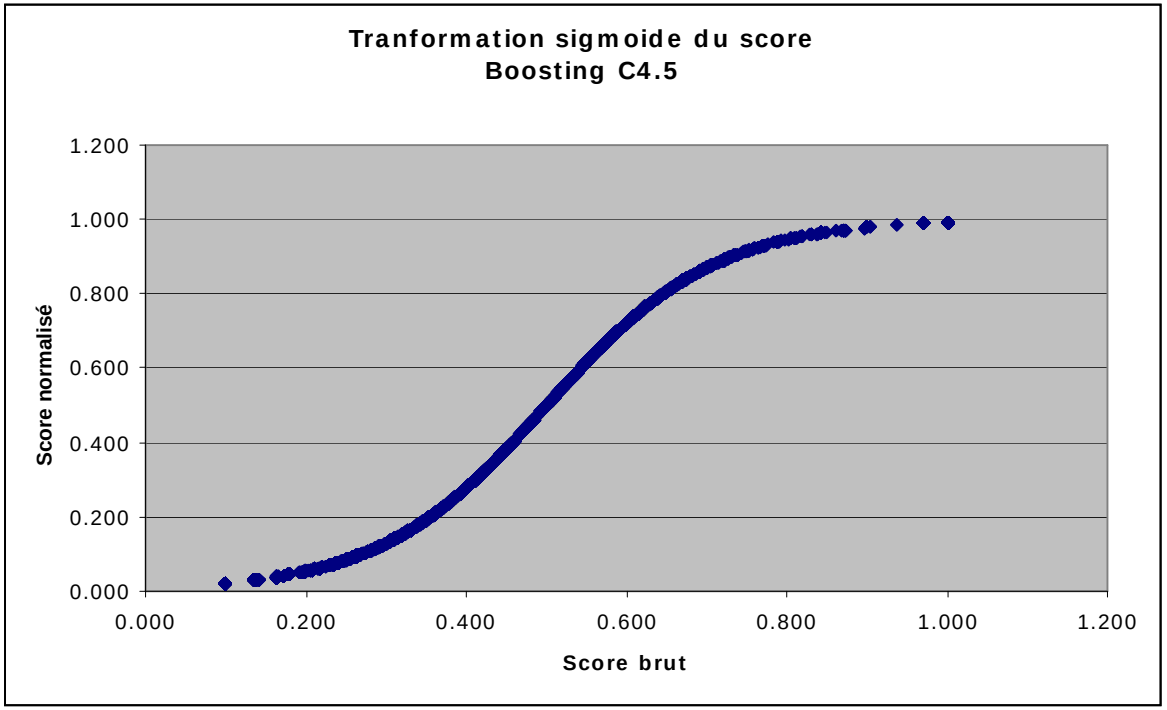
Démarche :
- Travailler sur le tuning set
- Estimer par maximum de vraisemblance les paramètres d'un fonction sigmoïde qui prend en entrée le score « brut » (s) et renvoie en « sortie » le score calibré (p)
- Pour chaque individu à « scorer », appliquer la transformation sur son score brut pour obtenir le score normalisé, comparable aux autres méthodes

Équation de transformation

$$p = \frac{1}{1+e^{a \times s + b}}$$

Estimé sur le « tuning » set

$$a = -9.53$$
$$b = 4.77$$



Normalisation des scores – Méthode de Platt (1999)

Utiliser la fonction sigmoïde – Détail des calculs

$$p_i = \frac{1}{1 + e^{a \times s_i + b}}$$

La (log)vraisemblance L à maximiser est égale à (de manière équivalente, on peut minimiser –L)

$$L = \sum_i y_i \times \ln(p_i) + (1 - y_i) \times \ln(1 - p_i)$$

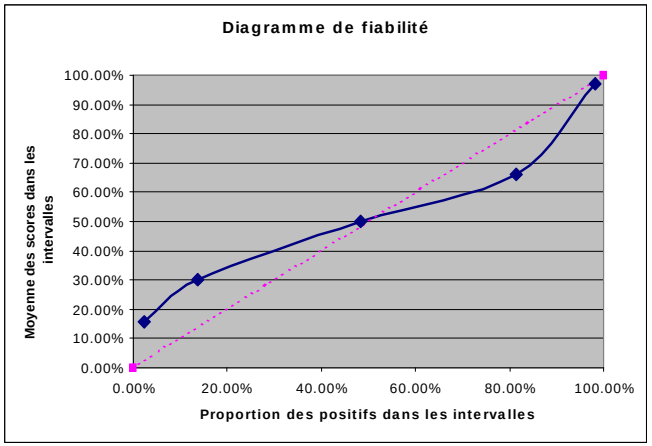
La maximisation (minimisation) s'appuie souvent sur une méthode d'optimisation numérique (ex. Newton-Raphson, etc.)

avec

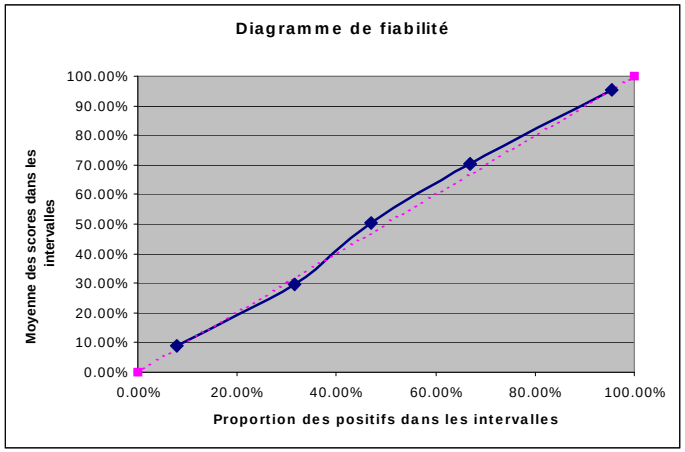
$$y_i = \begin{cases} \frac{n_+ + 1}{n_+ + 2}, \text{ si } y(i) = + \\ \frac{1}{n_- + 2}, \text{ si } y(i) = - \end{cases}$$

Utilisé à la place de y={0,1} pour relativiser l'appartenance à Y=+ ou Y=-. On introduit ainsi un « lissage » qui prévient le sur-apprentissage, surtout sensible lorsque le learning set sert également de tuning set.

Ex. Boosting C4.5
a = -1 et b=0 → -L = 320.03 / en utilisant le « solveur » d'Excel :a = -9.53, b = 4.77 → -L = 188



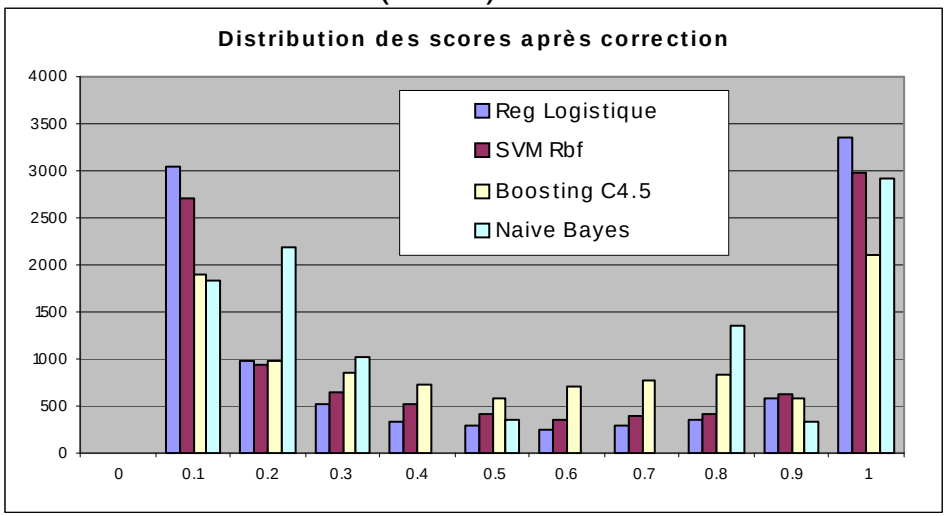
Avant normalisation, score brut…



Après, score calibré… presque aussi bien que la régression logistique !!!

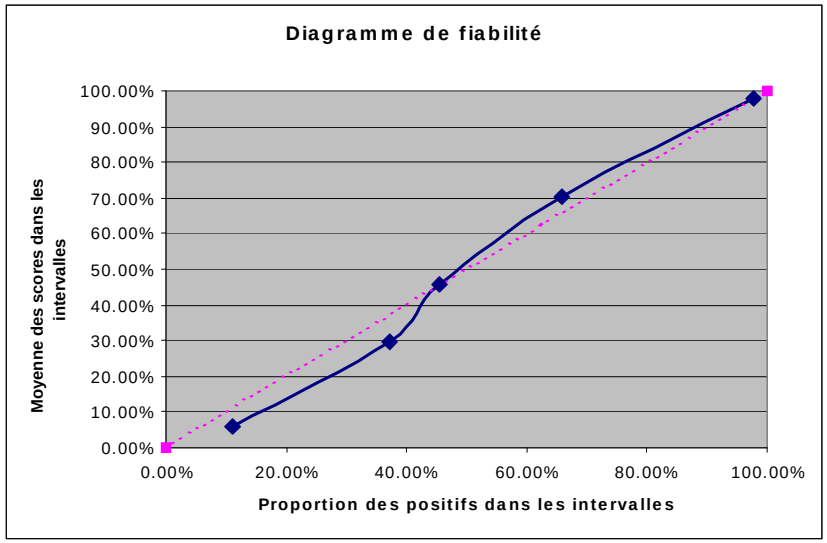
Normalisation des scores – Méthode de Platt (1999)

Quelques comparaisons

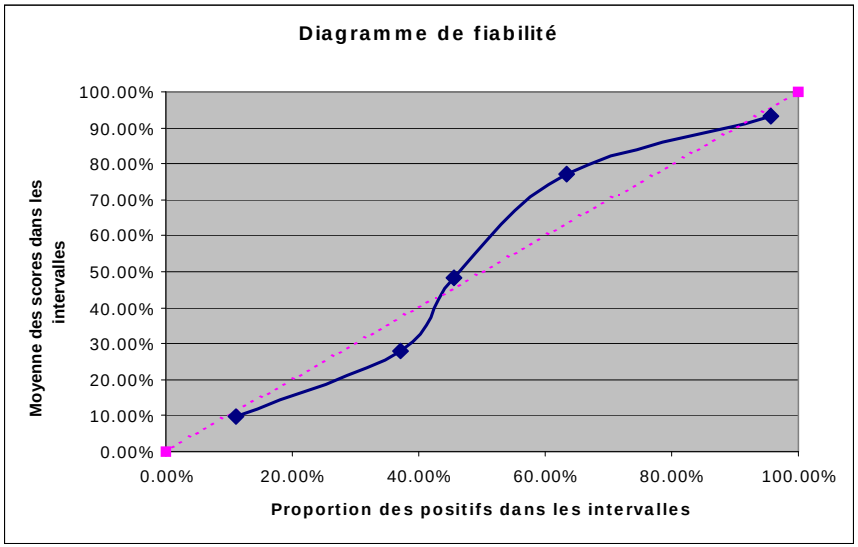


Tous les scores calibrés sont à peu près distribués de la même manière… maintenant ils sont comparables d'une méthode à l'autre, sauf pour Naive Bayes qui diffère significativement.

→ *En fait, elle a été mal normalisée par la méthode de Platt si on se réfère au diagramme de fiabilité… (la normalisation a même dégradé la situation initiale)*



Avant normalisation….



Après normalisation….

Normalisation des scores – La régression isotonique

Principe (Zadrozny & Elkan, 2001)

Idée : Construire une fonction de transformation « m » monotone croissante (isotonique)

Équation de régression : $y_i = m(s_i) + \varepsilon_i$ avec $y_i = \begin{cases} 1, \ si \ \ y(i) = + \\ 0, \ si \ \ y(i) = - \end{cases}$

Critère des moindres carrés : $\min \sum_i \left(y_i - m(s_i) \right)^2$

Deux éléments de différenciation avec l'approche précédente (Transf. Sigmoïde) :
- on optimise le critère des moindres carrés (et non plus la vraisemblance)
- la forme de la fonction « m » n'est pas spécifiée, on cherche parmi toutes les fonctions isotoniques possibles (*N.B. la fonction sigmoïde était isotonique d'ailleurs…*)

Démarche : Utilisation de l'algorithme PAV (Pool Adjacent Violators)

- Travailler sur le tuning set
- Calculer le score brut et trier les observations par score brut décroissant
- Rajouter une colonne des scores normalisés, la première observation a un score égal à 1, le dernier égal à 0, et par l'algorithme PAV, remplir le reste de la colonne
- Dans le fichier test et en déploiement, pour un nouvel individu à « scorer », calculer son score brut
- Puis, par interpolation, avec les deux colonnes scores bruts et scores normalisés du « tuning set », trouver la position du score brut et attribuer le score normalisé correspondant

Normalisation des scores – La régression isotonique

L'algorithme PAV (Pool Adjacent Violators)

cf. Fawcett & Niculescu-Mizil, 2006

Quelques idées maîtresses :

- le tableau est trié par score brut décroissant
- « initial » correspond à l'appartenance aux groupes
- cette colonne est transformée petit à petit en score normalisé
- il ne doit pas y avoir de « paires contradictoires » c.-à-d. un score normalisé plus petit placé devant un score normalisé plus grand (*sachant que le tableau a été trié de manière décroissante sur le score brut*)
- nous obtenons une fonction en escalier qui transforme le score brut en score normalisé
- on peut l'utiliser pour normaliser n'importe quel score brut en déploiement

Algorithm 1 Basic PAV method for generating probability estimates

Input: Scored training set (f_i, y_i), where f_i is the score assigned by the classifier and y_i is the correct class.

Output: Stepwise constant function generated by m

1: **begin**
2: Sort training set instances increasing by f_i
3: Put each training instance in its own group, $G_{i,i}$ and predict $m_{i,i} = y_i$ for it
4: **while** $\exists G_{k,i-1}$ and $G_{i,l}$ ST $m_{k,i-1} \geq m_{i,l}$ **do**
5: Pool the instances in $G_{k,i-1}$ and $G_{i,l}$ into one group, $G_{k,l}$
6: $m_{k,l} = (\sum_{i=k}^{l} y_i)/(l-k+1)$
7: Predict $m_{k,l}$ for all instances in $G_{k,l}$
8: **end while**
9: Output the stepwise constant function generated by m
10: **end**

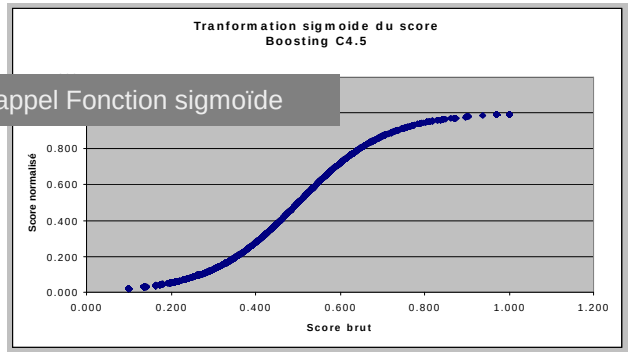
Table I. An illustration of the PAV algorithm

#	Score	Probabilities						
		Initial	**a1**	**a2**	**b**	**c1**	**c2**	**d**
0	.9	1	1	1	1	1	1	1
1	.8	1	1	1	1	1	1	1
2	.7	0	0	0	0	0	0	3/4
3	.6	1	1	1	1	1	1	3/4
4	.55	1	1	1	1	1	1	3/4
5	.5	1	1	1	1	1	1	3/4
6	.45	0	0	0	0	1/2	2/3	2/3
7	.4	1	1	1	1	1/2	2/3	2/3
8	.35	1	1	1	1	1	2/3	2/3
9	.3	0	0	0	1/2	1/2	1/2	1/2
10	.27	1	1	1	1/2	1/2	1/2	1/2
11	.2	0	0	1/3	1/3	1/3	1/3	1/3
12	.18	0	1/2	1/3	1/3	1/3	1/3	1/3
13	.1	1	1/2	1/3	1/3	1/3	1/3	1/3
14	.02	0	0	0	0	0	0	0

Normalisation des scores – Régression isotonique

Déploiement et évaluation (Boosting C4.5)

Fonction de transformation



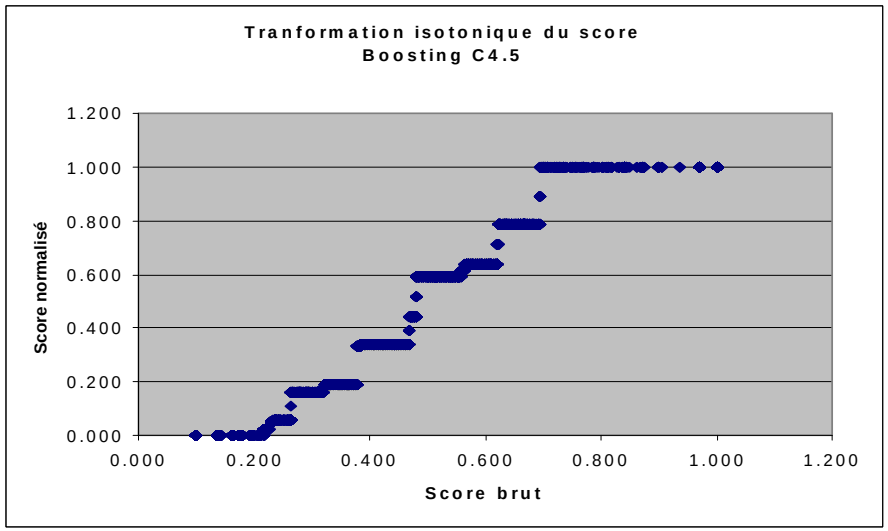
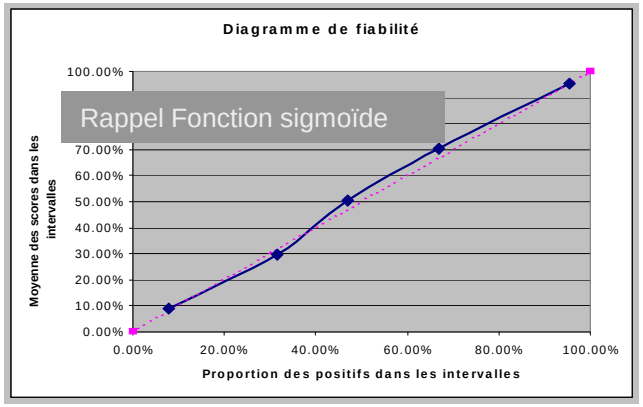
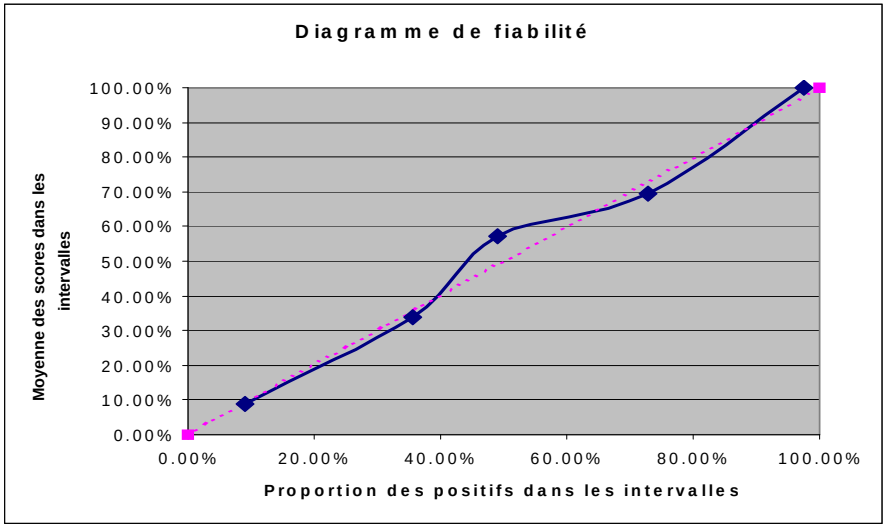


Diagramme de fiabilité





La transformation est bien en escalier, il produit des ex-æquo là où il n'y en avait pas sur les scores bruts (*c'est un inconvénient*), mais ça fonctionne bien (dixit le diagramme de fiabilité)

Normalisation des scores – Régression isotonique

Déploiement et évaluation (Naïve Bayes)

Fonction de transformation

Remarques :

• Il y a beaucoup d'ex-aequo, c'est inhérent à la technique

• C'est d'autant plus criant que Naïve Bayes produit elle aussi un score brut avec beaucoup d'ex-æquo (*cf. discrétisation, les observations dans les mêmes intervalles se voient attribuer le même score*)

• dans le cas présent, la fonction de transformation adéquate est linéaire et non sigmoïde, en s'affranchissant de la forme de la fonction, la méthode isotonique a pu la trouver !!!

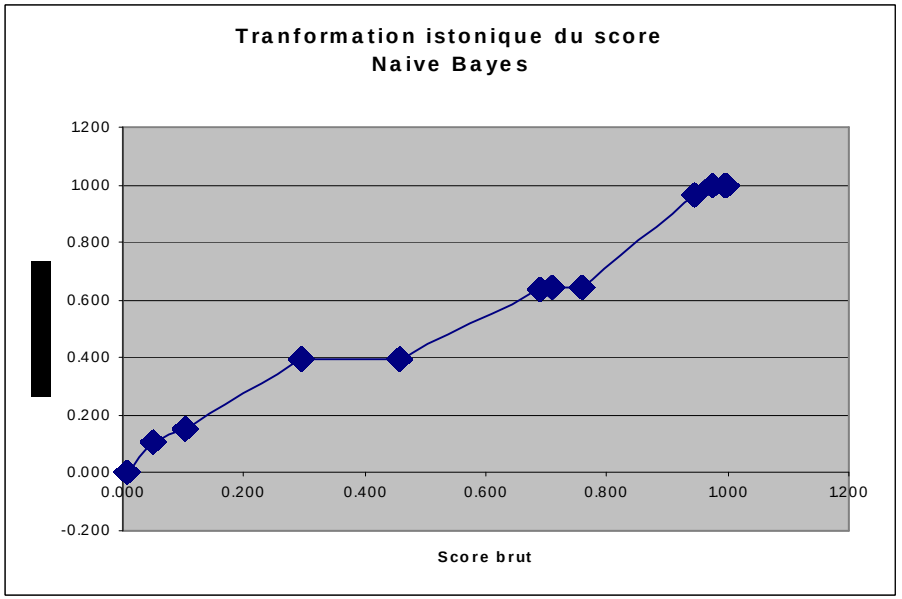
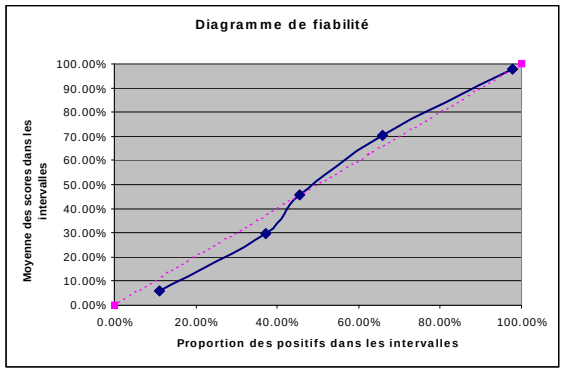
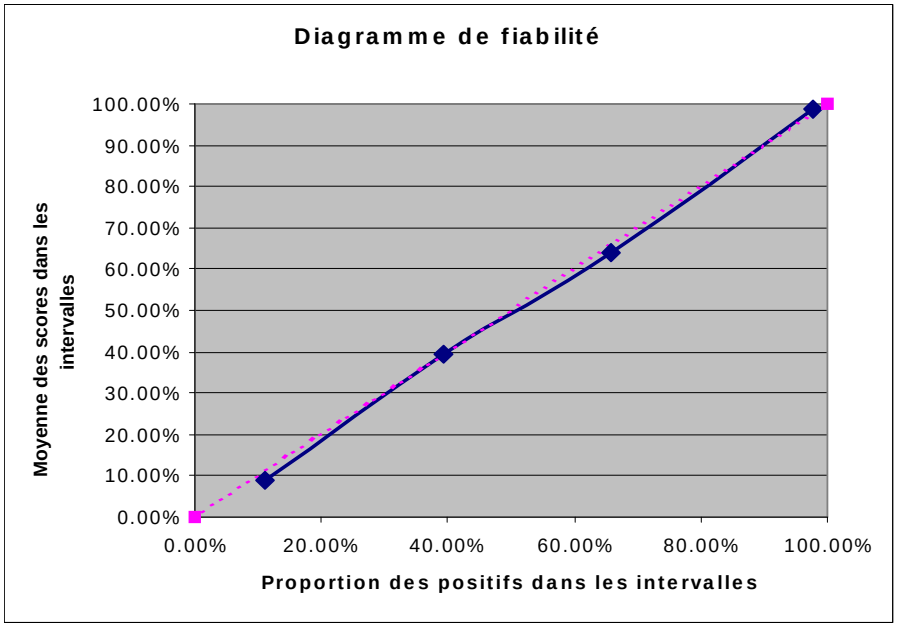


Diagramme de fiabilité

La qualité de transformation est remarquable !!!





Transformation sigmoïde
J. Platt, «Probabilistic outputs for support vector machines and comparison to regularized likelihood methods», Advances in Large Margin Classifiers, pp. 61-74, 1999.

Régression isotonique
B. Zadrozny, C. Elkan, «Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers », Proc. ICML, 2001.

Comparaison empirique et mise en œuvre avec plusieurs méthodes supervisées
A. Mizil, R. Caruana, «Predicting Good Probabilities with Supervised Learning », Proc. ICML, 2005.

Calibrer les scores est important dès qu'on veut les interpréter, appliquer des coûts de mauvaise affectation, comparer ou combiner des scores.

Certaines méthodes fournissent des scores naturellement bien calibrés et/ou se prêtent à des transformation analytiques simples et reconnues (cf. régression logistique, analyse discriminante, arbres de décision, bagging, etc.). D'autres méthodes, en revanche, sont mal calibrées (cf. SVM, Naive Bayes, Boosting, etc.)

Il y a deux techniques de transformation : la transformation sigmoïde (Platt, 1999), très simple mais pas toujours appropriée ; et la régression isotonique (Zadrozny & Elkan, 2001), plus générique mais lourde et compliquée à mettre en œuvre.