

Mesures de qualité des méthodes de classification

On se place dans le contexte de la classification binaire : $Y \in \{N, P\} = F$
et x vecteur de variables prédictives : $x \in E$.

A l'aide d'un échantillon d'apprentissage $(x_1, y_1), \dots, (x_n, y_n) = A_n$
on construit un classifieur $\hat{g}_{A_n}(\cdot) : x \mapsto \{N, P\}$

I - La matrice de confusion

Supposons que l'on dispose d'un échantillon de test à l'aide duquel nous allons évaluer la qualité du classifieur $\hat{g}_{A_n}(\cdot)$.

$$T_m = (\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_m, \tilde{y}_m)$$

Pour chaque \tilde{x}_j ($j=1, \dots, m$) on peut comparer \tilde{y}_j et $\hat{g}(\tilde{x}_j)$.

Le résultat de ces comparaisons peut être stocké dans une table de contingence qu'on appelle matrice de confusion : (taille 2×2)

Actual Prédictif	N	P
N	VN	FN
P	FP	VP

* VP : nb de Vrais Positifs

$$VP = \sum_{j=1}^m \underbrace{\mathbb{1}_{\{\tilde{y}_j = P\}}}_{\text{positifs}} \underbrace{\mathbb{1}_{\{\hat{g}_{A_n}(\tilde{x}_j) = P\}}}_{\text{prédits positifs}}$$

Il s'agit du nombre d'individus de test qui ont été bien classés parmi les positifs (vrais)

* FP : Faux Positifs : $FP = \sum_{j=1}^m \mathbb{1}_{\{\tilde{y}_j = N\}} \mathbb{1}_{\{\hat{g}_{A_n}(\tilde{x}_j) = P\}}$

C'est le nombre de mal classés parmi ceux prédits positifs.

* VN Vrais Négatifs : $VN = \sum_{j=1}^m \mathbb{1}_{\{\tilde{y}_j = N\}} \mathbb{1}_{\{\hat{y}_m(\tilde{x}_j) = N\}}$

= nb de bien classés parmi les négatifs.

* FN Faux Négatifs : $FN = \sum_{j=1}^m \mathbb{1}_{\{\tilde{y}_j = L\}} \mathbb{1}_{\{\hat{y}_m(\tilde{x}_j) = N\}}$

= nb de mal classés parmi ceux classés négatifs.

⇒ On a bien sur $m = VP + FN + FP + VN$

On définit à présent plusieurs notions :

* Accuracy (Précision) = pourcentage de bien classés
 $= \frac{VP + VN}{m}$

* Error rate (Taux d'erreur) = pourcentage de mal classés
 $= 1 - \frac{VP + VN}{m} = \frac{FP + FN}{m}$

* Sensitivity (Sensibilité) = pourcentage de bien classés parmi les positifs
 $= \frac{VP}{VP + FN} = \frac{VP}{M_P}$
 (nb de positifs de mon éch. test)

* Specificity (Spécificité) = pourcentage de bien classés parmi les négatifs
 $= \frac{VN}{VN + FP} = \frac{VN}{M_N}$
 (nb de Négatifs de mon échantillon test)

On en déduit alors :

$$\text{Accuracy} = \frac{M_P}{m} \cdot \text{Sensitivity} + \frac{M_N}{m} \cdot \text{Specificity}$$

Maximiser l'accuracy \Leftrightarrow minimiser l'error rate

↳ cela correspond à une fonction de coût élémentaire symétrique

$$\text{égal à } 1 : h(P, N) = h(N, P) = 1,$$

pour laquelle le classifieur optimal est le classifieur de Bayes, celui qui affecte à la classe de probabilité la plus forte :

$$g^*(x) = \begin{cases} P & \text{si } P(Y=P | X=x) \geq 1/2 \\ N & \text{si } P(Y=P | X=x) < 1/2 \end{cases}$$

II - Courbes ROC (Receiving Operator Characteristic)

Les courbes sont utilisées pour comparer des techniques de classification basées sur des scores.

$$S_1(x) = \hat{P}^{M_1} (Y=P | X=x)$$

$$S_2(x) = \hat{P}^{M_2} (Y=P | X=x)$$

M_1 et M_2 sont deux méthodes de classification.

Par exemple, $M_1 = \text{LDA}$ et $M_2 = \text{régression logistique}$

A partir de ces scores, on peut définir des classifieurs (plusieurs par scores)

Soit λ un seuil $\in [0, 1]$

$$\hat{g}_\lambda^{M_1}(x) = \begin{cases} P & \text{si } S_1(x) \geq \lambda \\ N & \text{sinon} \end{cases}$$

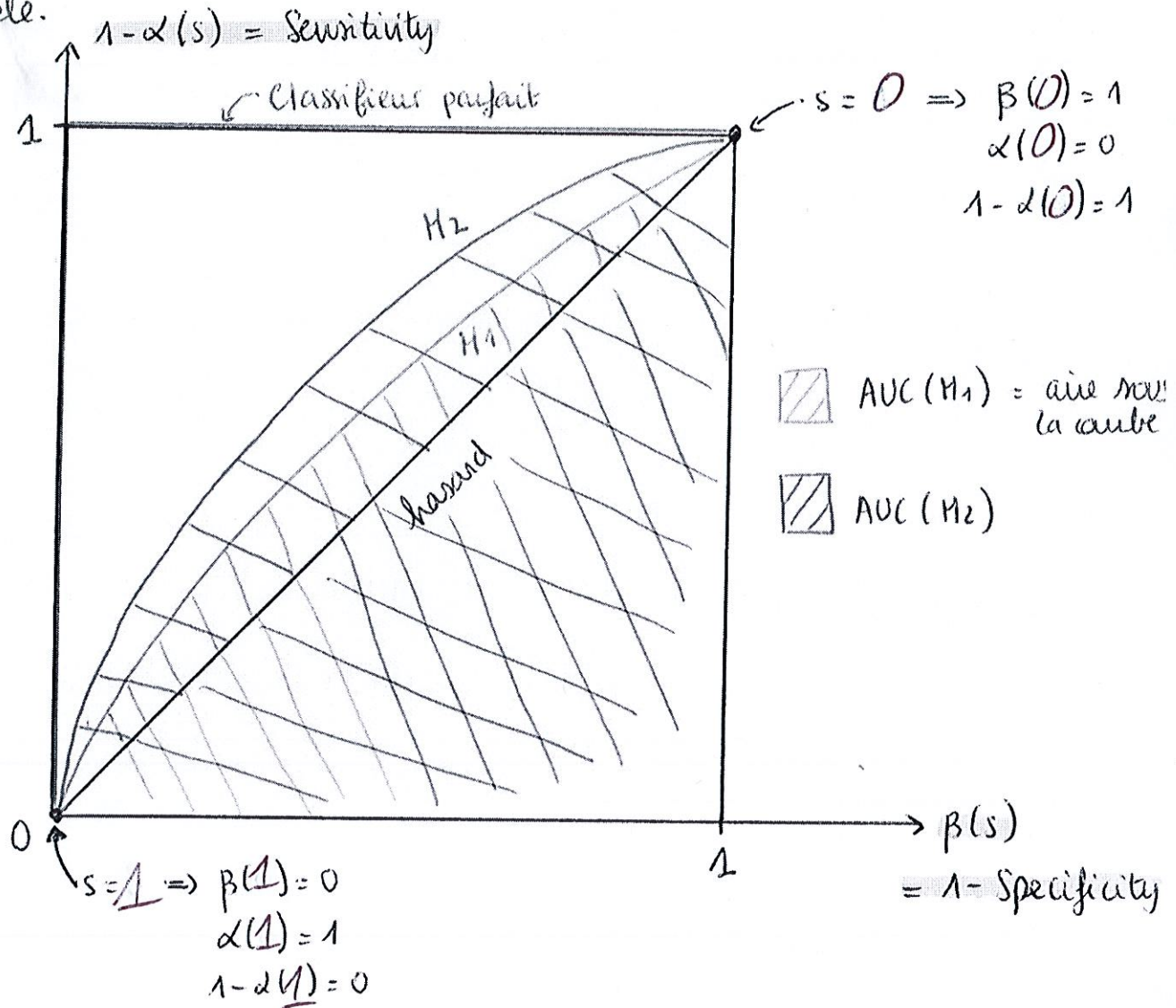
$$\hat{g}_\lambda^{M_2}(x) = \begin{cases} P & \text{si } S_2(x) \geq \lambda \\ N & \text{sinon} \end{cases}$$

$$\beta(\lambda) = P(S_1(x) \geq \lambda | Y=0) = \frac{FP}{N} = 1 - \text{Specificity}$$

$$\alpha(\lambda) = P(S_1(x) < \lambda | Y=1) = \frac{FN}{P} = 1 - \text{Sensitivity}$$

La courbe ROC est une courbe paramétrée ayant en abscisse $\beta(s)$ (21) et en ordonnée $1 - \alpha(s)$.

Pour s donné, plus $\beta(s)$ est faible et $1 - \alpha(s)$ est fort, meilleur est le modèle.



Plus on s'éloigne de la diagonale, meilleur est la méthode de la classification. On préférera la technique telle que l'aire sous la courbe ROC est la plus importante (AUC le plus grand).

Pour déterminer un classifieur, il faut fixer la valeur de s .

Pour la méthode qui a l'AUC la + forte, on prendra la valeur de s telle que la courbe ROC est la plus éloignée possible de la première bissectrice.

III - Kappa de Cohen

Le coefficient Kappa compare la qualité d'un classifieur à une concordance électeur obtenue en utilisant les lois marginales de la matrice de confusion -

$$Kappa = \left[\frac{Accuracy - \text{random Accuracy}}{1 - \text{random Accuracy}} \right]$$

Attention = random Accuracy

≠ Accuracy of random classifier

En effet, il y a deux types de classifieur électeur -
1] le classifieur électeur non pondéré

$$g^u(\pi) = \begin{cases} N \text{ avec proba. } \frac{1}{2} \\ P \text{ avec proba. } \frac{1}{2} \end{cases}$$

(3)

le terme u / v de la classification est égal à $\frac{1}{2} \Rightarrow$ son accuracy

est égal à $\frac{1}{2}$

ij la classification ultérieure possible

$$g^{u,v}(\pi) = \begin{cases} N \text{ avec proba. } \frac{MN}{M} \\ P \text{ avec proba. } \frac{MP}{M} \end{cases}$$

son accuracy est

$$= \frac{\frac{MN}{M} \frac{MN}{M}}{M^2} + \frac{\frac{MP}{M} \frac{MP}{M}}{M^2} = \frac{1 - 2MN + 2M^2}{M^2}$$

Ces deux termes ne correspondent pas à la définition de la Random Accuracy

Scrit la matrice de confuzion (4)

Obs.

Real. \ Obs.	N	P	
N	VN	FN	M'_N
P	FR	VR	M'_P
	M_N	M_P	

$$\text{Precision Accuracy} = \frac{M'_N M_N}{M^2} + \frac{M'_P M_P}{M^2}$$

Sum

Real. \ Obs.	N	P	
N	0.1	0.3	0.4
P	0.1	0.5	0.6
	0.2	0.8	

meas of errors

$$\begin{aligned} \text{Precision Accuracy} &= 0.2 \times 0.4 + 0.8 \times 0.6 \\ &= 0.56 \end{aligned}$$