Classification supervisée
Evaluation d'une règle de classement

V. Vandewalle (vincent.vandewalle@univ-lille.fr)

Polytech'Lille GIS2A4

Année universitaire 2019-2020

Règle d'affectation et probabilité d'erreur - Définitions

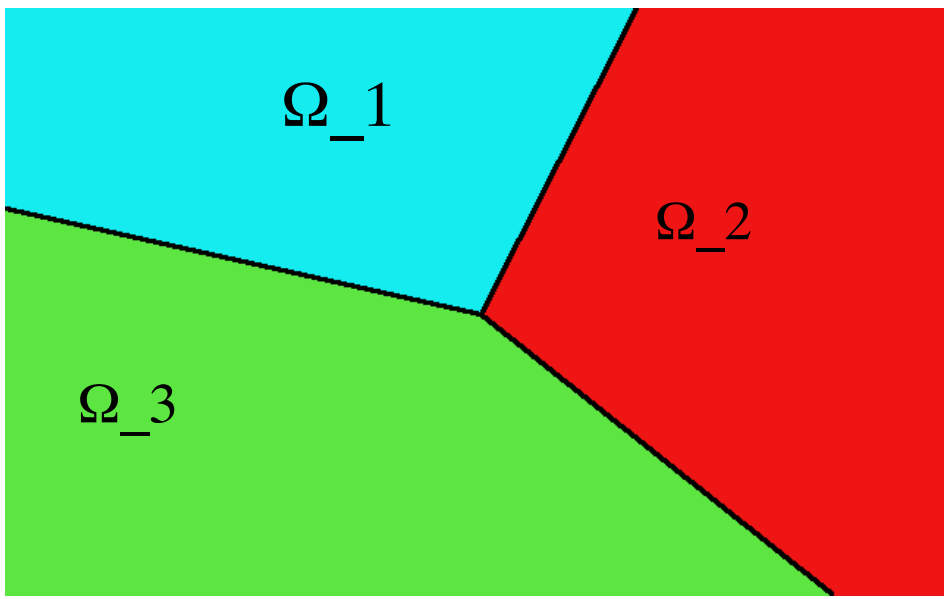
Rappel :

Règle d'affectation (ou de classement, de décision, ...) :

$$r : x \in \mathbb{R}^p \mapsto \hat{y} = r(x) \in \{1, 2, \ldots, K\}$$

Définir r revient à **partitionner** $\mathbb{R}^p$ en K régions Ω_k telles que

$$x \in \Omega_k \Leftrightarrow r(x) = k$$



Règle d'affectation et probabilité d'erreur - Définitions

Probabilité de classer un individu de G_k dans G_ℓ avec la règle r :

$$e_{k\ell}(r) = P(r(X) = \ell | Y = k) = \int_{\Omega_\ell} f_k(x)dx.$$

Probabilité qu'un individu de G_i soit mal classé avec la règle r :

$$e_i(r) = P(r(X) \neq i | Y = i) = \sum_{\ell \neq i} e_{i\ell}(r) = \int_{\bar{\Omega}_i} f_i(x)dx.$$

avec $\bar{\Omega}_i$ le complémentaire de Ω_i

Probabilité de mauvais classement global (ou erreur globale de classement) :

$$e(r) = \sum_{i=1}^{K} \pi_i e_i(r) \quad \text{avec } \pi_i = P(Y = i)$$

Règle de classement optimale

Objectif : Définir la meilleure règle de classement possible.

Définition :

Coût de mauvais classement de classer un individu de G_ℓ dans G_k

$$C : (k, \ell) \in \{1, \ldots, K\} \times \{1, \ldots, K\} \to C(k, \ell) \in \mathbb{R}^+,$$

où par convention $C(k, k) = 0$.

Les fonctions de coût ne sont généralement pas symétriques : classer un individu sain comme malade n'a pas le même coût que l'erreur inverse.

Règle de classement optimale

Coûts à définir :

- avec le praticien en fonction de son expérience,
- en testant plusieurs systèmes de coûts possibles et en comparant les résultats obtenus,
- en les fixant tous à 1 lorsque l'on a aucune idée.

Règle de classement optimale

Définitions

Risque conditionnel associé à $x \Rightarrow$ coût moyen de classement :

$$R(r(X)|X = x) = E[C(r(X), Y)|X = x] = \sum_{i=1}^{K} C(r(x), i)t_i(x),$$

avec $t_i(x) = P(Y = i|X = x) = \dfrac{\pi_i f_i(x)}{f_X(x)} = \dfrac{\pi_i f_i(x)}{\sum_{i'=1}^{K} \pi_{i'} f_{i'}(x)}$

(probabilité conditionnelle)

Risque moyen $\Rightarrow$ coût moyen de classement inconditionnel

$$R(r) = E_X[R(r(X)|X = x)] = \sum_{i=1}^{K} \pi_i \sum_{\ell=1}^{K} C(\ell, i) \int_{\Omega_\ell} f_i(x)dx.$$

Règle de classement optimale

Règle de classement optimale r^* : règle qui minimise le risque moyen $\Rightarrow$ cela revient à minimiser le risque conditionnel pour chaque individu car :

$$R(r^*) = \min_r E_X[R(r(X)|X = x)] \geq E_X[\min_r R(r(X)|X = x)].$$

La règle optimale affecte donc x à G_i si

$$R(r(X) = i|X = x) < R(r(X) = \ell|X = x) \qquad \forall \ell \neq i.$$

Règle de classement optimale

$$R(r(X) = i|X = x) = E[C(i, Y)|X = x] = \sum_{\ell=1}^{K} C(i, \ell)t_\ell(x),$$

Règle optimale de Bayes :

$$r^*(x) = i \qquad \text{si } \sum_{\ell=1}^{K} C(i, \ell)t_\ell(x) < \sum_{\ell=1}^{K} C(k, \ell)t_\ell(x) \quad \forall k \neq i.$$

Par exemple pour $K = 3$

$$\begin{pmatrix} C(1,1) & C(1,2) & C(1,3) \\ C(2,1) & C(2,2) & C(2,3) \\ C(3,1) & C(3,2) & C(3,3) \end{pmatrix} \begin{pmatrix} t_1(x) \\ t_2(x) \\ t_3(x) \end{pmatrix} = \begin{pmatrix} R(r(X) = 1|X = x) \\ R(r(X) = 2|X = x) \\ R(r(X) = 3|X = x) \end{pmatrix}$$

On affecte l'individu x à la classe dans laquelle le risque conditionnel $R(r(X) = i|X = x)$ est le plus faible.

Règle de classement optimale

Cas de l'égalité des coûts

Si tous les coûts sont égaux à c, le risque conditionnel est alors

$$R(r(X) = i|X = x) = c \sum_{k \neq i}^{K} t_k(x) = c(1 - t_i(x)),$$

et donc $r^*(x) = i$ si $c(1 - t_i(x)) < c(1 - t_k(x)) \quad \forall k \neq i$ ou encore

$$r^*(x) = i \qquad \text{si } t_i(x) > t_k(x) \quad \forall k \neq i.$$

L'observation x est donc affectée à la classe conditionnellement la plus probable (règle du *maximum a posteriori : MAP*).

Règle de classement optimale

Les coûts étant égaux, en posant $c = 1$, le risque moyen de classement

$$\begin{aligned} R(r) &= \sum_{i=1}^{K} \pi_i \sum_{l \neq i} \int_{\Omega_l} f_i(x) dx \\ &= \sum_{i=1}^{K} \pi_i \int_{\bar{\Omega}_i} f_i(x) dx \\ &= \sum_{i=1}^{K} \pi_i e_i(r) \\ &= e(r) \end{aligned}$$

est égal à l'erreur globale de classement.

Règle de classement optimale

Cas de deux classes

On a

$$\begin{aligned} & r^*(x) = 1 \quad \text{si} \quad C(1,2)t_2(x) < C(2,1)t_1(x), \\ \text{et} \quad & r^*(x) = 2 \quad \text{si} \quad C(2,1)t_1(x) < C(1,2)t_2(x), \end{aligned}$$

soit en posant $g(x) = \frac{C(2,1)t_1(x)}{C(1,2)t_2(x)}$

$$\begin{aligned} & r^*(x) = 1 \quad \text{si} \quad g(x) > 1, \\ \text{et} \quad & r^*(x) = 2 \quad \text{si} \quad g(x) < 1. \end{aligned}$$

L'équation de la *surface discriminante* (ou *frontière de classement*) est $g(x) = 1$.

Estimation du taux d'erreur

Erreur apparente $\hat{e}^a$

Règle de classement appliquée sur les observations qui ont servi à apprendre le modèle.

On montre que cet estimateur est en général biaisé et optimiste. Cet estimateur est donc à éviter.

Il donne cependant une minoration de l'erreur de classement réelle.

Par exemple si $\hat{e}^a = 35\%$, alors en réalité l'erreur de classement devrait être supérieure à 35%.

Estimation du taux d'erreur

Méthode de la partition $\hat{e}^p$

Division de l'échantillon en un **échantillon d'apprentissage** (environ 2/3 à 75% de l'échantillon global) et un **échantillon test**. L'erreur de classement pourra alors être estimée sans biais sur l'échantillon test.

Remarque : Cette technique demande une taille d'échantillon suffisamment grande.

Estimation du taux d'erreur

Méthode de la validation croisée

Estimateur **validation croisée leave-one-out** de l'erreur :

$$\hat{e}^{cv} = \frac{1}{n} \sum_{h=1}^{n} \hat{e}^{p}_{(h)}$$

où $\hat{e}^{p}_{(h)}$ est l'évaluation de l'erreur sur une partition test constituée d'uniquement l'observation h : $(x, y)_h$.

On parle de **validation croisée** v**-fold** lorsque l'échantillon initial est partagé en v sous-échantillons servant chacun tour à tour d'échantillon test tandis que le reste des échantillons est utilisé pour l'apprentissage.

Estimation du taux d'erreur

On montre que l'on obtient un **estimateur sans biais de l'erreur**, ayant une **variance plus faible que** $\hat{e}^p$ avec une partition test réduite à une seule observation.

Remarques :

- Cette technique demande de ré-estimer les paramètres pour chaque échantillon test considéré. Dans le cas de la validation croisée leave-one-out, les paramètres du modèle sont donc estimés n fois.
- Cette technique est à privilégier dans le cas de petits échantillons.

Evaluation du modèle final

Si plusieurs modèles sont évalués sur un échantillon test et que le meilleur est retenu, il y a un risque de sur-estimation des performances de ce modèle.

Dans ce cas, on utilise un **échantillon de validation** ($\approx 10\%$) pour estimer sans biais les **performances de ce modèle final**.

Evaluation du modèle final

Cas binaire ($K = 2$)

- si les coûts sont égaux, règle du MAP
- si les coûts sont différents, on peut avoir un seuil s différent de 0.5

$$\hat{y_s} = \begin{cases} 1 \text{ si } P(Y = 1/X = x^*) \geq s \\ 0 \text{ sinon} \end{cases}$$

Evaluation du pouvoir discriminant (= capacité du modèle à correctement classer les observations) :

- **matrice de confusion** (dépend du seuil)
- **courbe ROC** (ne dépend pas d'un seuil)

Evaluation du modèle final

Matrice de confusion (à seuil s fixé)

	y prédits $= 0$	y prédits $= 1$
y réels $= 0$	Vrais Négatifs (VN)	Faux Positifs (FP)
y réels $= 1$	Faux Négatifs (FN)	Vrais Positifs (VP)

Taux de bon classement :

$$TBC = \frac{VN + VP}{n}$$

Evaluation du modèle final

Sensibilité (% de vrais positifs) :

$$Se = \frac{VP}{VP + FN}$$

Spécificité (% de vrais négatifs) :

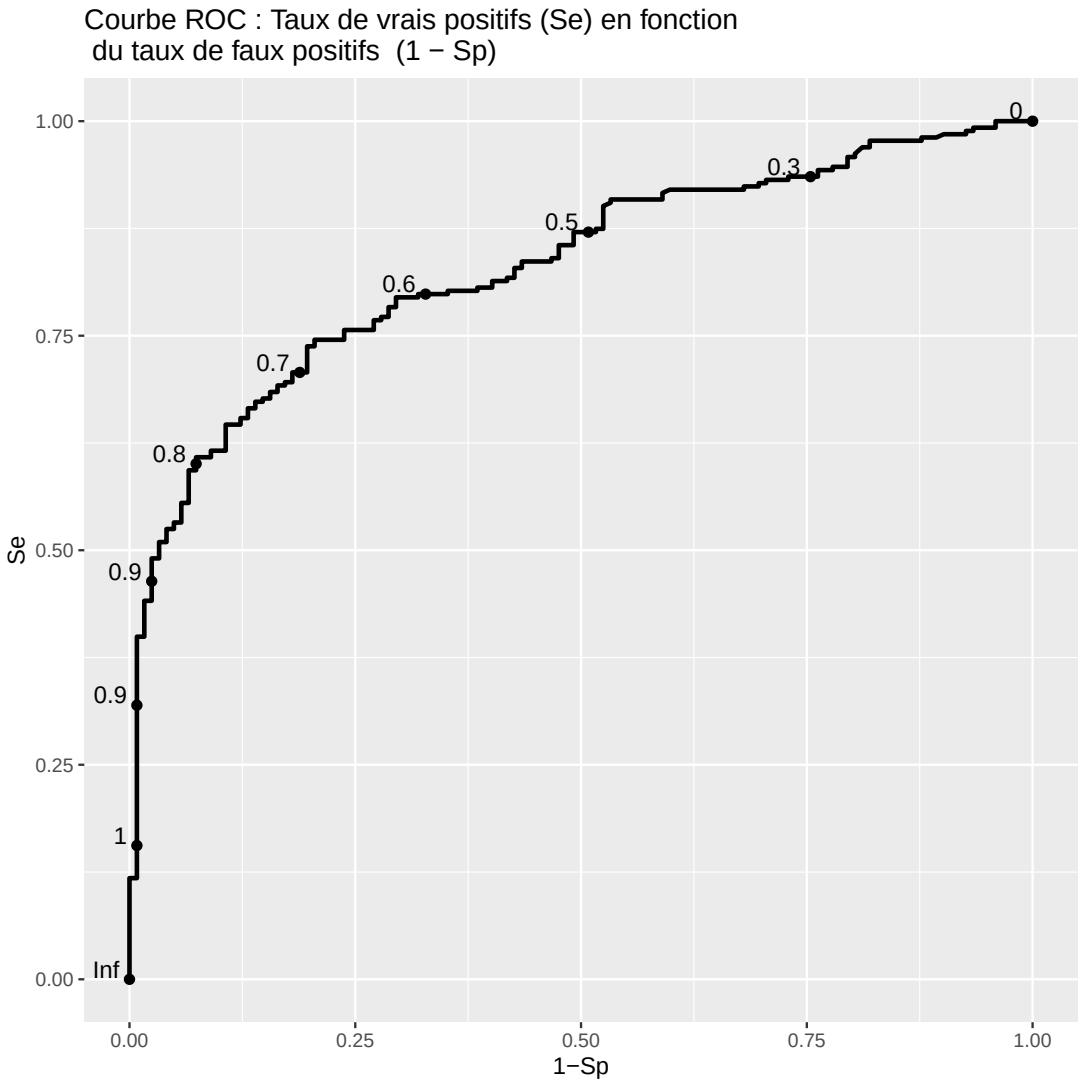
$$Sp = \frac{VN}{VN + FP} = 1 - \frac{FP}{VN + FP}$$

Point idéal : $Se = 1$, $Sp = 1$ (dans la réalité, jamais atteint)

Courbe ROC : courbe paramétrée par s qui trace Se en fonction de $(1 - Sp)$

AUC : aire sous la courbe ROC (si égale à 0,5 quand décision due au hasard ...)

Exemple de courbe ROC



Évaluation du modèle final

La courbe ROC permet de choisir le seuil

- par contrainte métier (maximiser Se ou Sp prioritairement)
- en cherchant le point le plus proche du point idéal :

$$\underset{s}{\operatorname{argmin}} \ (1 - Se(s))^2 + (1 - Sp(s))^2$$

Remarque : On trace parfois le *rappel (Recall)* en fonction de la *précision (Precision)*

Dans ce cas,

- rappel = sensibilité
- précision $= \dfrac{VP}{VP + FP} = 1 - \dfrac{FP}{VP + FP}$