PROBLÈME DE CLASSIFICATION (I7)
(26 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **problème de classification** est un **problème de décision** statistique dans lequel l'**ensemble** des **décisions** possède un nombre fini d'éléments.

Plus précisément, on cherche à décider si une **unité statistique** donnée appartient à un ensemble (eg une **population statistique**) donné(e) parmi divers ensembles possibles.

En pratique, ceci est réalisé à partir d'une « **mesure** », ou **observation**, effectuée sur cette unité : en considérant cette mesure comme une **variable aléatoire**, il est équivalent de tester si celle-ci est générée par une **loi de probabilité** parmi plusieurs lp (cf **classification**).

Autrement dit, on cherche à confirmer un **hypothèse**, parmi plusieurs, à l'aide d'une observation donnée.

(i) Soit $(\Omega_i , \mathcal{T}_i)_{i=1,...,k}$ une **famille** d'**espaces probabilisables**, $(\mathcal{X}, \mathcal{B})$ un **espace d'observation** et $\xi_i : \Omega_i \mapsto \mathcal{X}$ une suite de va donnée ($\forall$ i $\in$ N_k*). On note $\Pi_{\mathcal{X}}$ = {w_1 ,..., w_k} une **partition** (mesurable) de $\mathcal{X}$, tq $w_i = \xi_i (\Omega_i)$, $\forall$ i $\in$ N_k*. Dans ce cadre :

(a) une **règle de décision pure**, ou **règle de décision déterministe**, consiste à imputer une unité $\omega = \xi_i^{-1} (x)$ à la population $\xi_i^{-1} (w_i) = \Omega_i$ ssi il existe un indice i $\in$ N_k* tq x $\in$ w_i ;

(b) une **règle de décision mixte**, ou **règle de décision aléatoire**, est une fonction (vectorielle) m : x $\mapsto$ m (x), à valeurs dans le **simplexe** S_k de $\mathbf{R}^k$, qui associe à l'observation x $\in$ $\mathcal{X}$ (resp à l'unité $\omega \in \xi_i^{-1} (x)$) la classe $w_i \subset \mathcal{X}$ (resp la population $\Omega_i = \xi_i^{-1} (w_i)$) avec la **probabilité** $m_i (x)$. Autrement dit, après observation de x, on tire un **nombre au hasard** dans N_k* selon la probabilité dont la répartition des masses est {$m_1 (x)$,..., $m_k (x)$} : si le résultat est l'**indice** i, on décide alors que x $\in$ w_i (resp que $\omega \in \xi_i^{-1} (x)$).

Soit, $\forall$ i $\in$ N_k*, une **mesure positive** μ_i et une **mesure de probabilité** P_i définie sur $\mathcal{T}_i$. On suppose que μ_i est une **mesure σ-finie** sur $\mathcal{B}$ et que l'image $P^{\xi(i)} = \xi_i (P_i)$ de P_i admet la **densité** (ou **dérivée de NIKODYM-RADON**) $f_i = dP^{\xi(i)} / d\mu_i$ pr à μ_i .

On définit une **matrice** $\Lambda = (\lambda_{ij})_{(i , j)} \in M_k (\mathbf{R})$, appelée **matrice de perte**, dont le terme général λ_{ij} représente la perte résultant de la décision $\omega \in \Omega_j$ alors que $\omega \in \Omega_i$ (cf **fonction de perte**). On suppose usuellement que la diagonale de Λ est nulle, ie λ_{ii} = 0, $\forall$ i $\in$ N_k*.

La **perte moyenne** dûe aux décisions $\omega \in \Omega_i$ $(\forall\, j \in N_k^*)$ alors que $\omega \in \Omega_i$ s'écrit :

(1)
$$L_i = \sum_{j=1}^{k} \int_{w(j)} \lambda_{ij}\, dP^{\xi(i)} \qquad \text{(dans le cas d'une règle pure)},$$
$$L_i = \sum_{j=1}^{k} \int_{\mathcal{X}} \lambda_{ij}\, m_j\, dP^{\xi(i)} \qquad \text{(dans le cas d'une règle mixte)},$$

où x(i) désigne x_i et w(j) désigne w_j $(\forall\, j \in N_k^*)$.

L'ensemble $\mathscr{L}$ des fonctions de perte vectorielles (en anglais « *operating caracteristics* ») $L = (L_1, ..., L_k)'$ ainsi définies est muni d'une **relation d'ordre** partiel $\prec\!\prec$ selon :

(2) $\quad L' \prec\!\prec L'' \Leftrightarrow L_i' \le L_i'', \qquad \forall\, i \in N_k^*,$

l'ordre strict correspondant étant :

(3) $\quad L' \prec L'' \Leftrightarrow L' \prec\!\prec L''$ et $\exists\, i \in N_k^*$ tq $L_i' < L_i''$.

(ii) On appelle **(règle de) décision admissible** toute règle (pure ou mixte, selon le cas) dont la fonction de perte vectorielle $L^* = (L_1^*, ..., L_k^*)'$ est un élément extrémal de l'ensemble préordonné $(\mathscr{L}, \prec\!\prec)$ (cf **préordre**). L'ensemble $\mathscr{A}$ des règles admissibles (règles « candidates ») comporte, en général, trop d'éléments parmi lesquels décider.

Divers **principes de réduction** sont alors utilisés, notamment le principe du « minimax » (défini à partir de **règles « minimax »**), ou le **principe bayésien** (défini à partir de **règles de BAYES**). Ainsi :

(a) on appelle **(règle de) décision « minimax »** la règle $\delta^* \in \mathscr{A}$ dont la fonction de perte L^* vérifie :

(4) $\quad \max_{i=1}^{k} L_i^* = \min_{\delta \in \mathscr{A}} \max_{i=1}^{k} L_i$.

(b) si l'on connaît la **probabilité a priori** Π_i de la population Ω_i (avec $i \in N_k^*$ et $\Pi = (\Pi_1, ..., \Pi_k) \in S_k$ (**simplexe** de $\mathbf{R}^k$)), ie si l'on mélange les populations Ω_i (resp les probabilités P_i) à l'aide de Π (cf **mélange légal**), on se ramène à une perte moyenne scalaire :

(5) $\quad l = \sum_{i=1}^{k} \Pi_i L_i$.

Dans le cas d'une règle pure, on appelle **indice de discrimination**, ou **« score » discriminant**, associé à Ω_i la quantité :

(6) $\quad S_i = -\sum_{j=1}^{k} \Pi_j \lambda_{ji} f_i$, $\qquad \forall\, i \in N_k^*$.

Par suite, la **perte moyenne** s'écrit :

(7) $I = - \Sigma_{i=1}^{k} \int_{w(j)} S_i \, d\mu_i$.

Dans le cas d'une règle mixte, on obtient :

(8) $I = - \Sigma_{i=1}^{k} \int_{\mathcal{X}} m_i \, S_i \, d\mu_i$.

(iii) Par suite, une solution bayésienne optimale consiste à décider $\omega \in \Omega_i$ si la probabilité a posteriori de Ω_i , ie :

(9) $\beta_i = \Pi_i f_i \, / \Sigma_{j=1}^{k} \Pi_j f_j$,

est maximum, ie si $\beta_j < \beta_i$, $\forall \, j \neq i$.

(iv) A titre d'exemple, si $P^{\xi(i)} = \mathcal{N}_Q (\mu_i , \Sigma_i)$ (**loi normale**, où μ_i désigne l'**espérance** de ξ_i), on obtient un **indice de discrimination quadratique** :

(10) $S_i = - (1/2) . \text{Log} |\Sigma_i| - (1/2) . (x_i - \mu_i)' \, \Sigma_i^{-1} (x_i - \mu_i) + \text{Log} \, \Pi_i$.

(v) Soit $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ un **modèle statistique** dans lequel la **famille** $\mathcal{P}^X$ est partitionnée en k sous-familles $\mathcal{P}_1^X$,..., $\mathcal{P}_k^X$ susceptibles de générer une observation donnée $x = X (\omega) \in \mathcal{X}$.

On appelle **problème de classification**, ou **problème à éventualités multiples**, ou encore **problème de (ou à) décisions multiples**, le problème du choix de la sous-famille $\mathcal{P}_i^X$ ($i \in N_k^*$) qui est à l'origine de x (cf aussi **analyse discriminante**, **classification**).

(vi) Un problème de classification constitue ainsi un exemple important de **problème de décision multiple**, dans lequel :

(a) l'espace des décisions (ou actions) $(D, \mathcal{B}_D)$ est fini :

(11) $D = \{d_1 ,..., d_k\}$;

(b) l'espace des paramètres $(\Theta, \mathcal{B}_\Theta)$ est fini et a même cardinalité que D, ie :

(12) $\Theta = \{\theta_1 ,..., \theta_k\}$;

(c) le nombre k des décisions possibles vérifie $k \geq 3$;

(d) on observe une va X : $\Omega \mapsto \mathcal{X}$ dont une des lois possibles $P_{\theta(i)}^{X}$ est généralement dominée par une mesure positive σ-finie μ définie sur $\mathcal{B}$ (**tribu de parties** de $\mathcal{X}$), ie est une **loi à densité** :

(13) $dP_{\theta(i)}^{X}(x) = f(x, \theta_i)\, d\mu(x)$.

où les $\theta(i)$ désignent par commodité les θ_i .

Dans ce cadre, pour classer X, ie pour déterminer la loi $P_{\theta(i)}^{X}$ qui a généré X (ie la lp suivie par X), on définit, conformément à la **théorie de la décision** statistique, une fonction de perte, eg de la forme :

(14) $L(\theta_i , d_j) =$ 1 si $j \neq i$ (classification incorrecte), 0 si $j = i$ (classification correcte),

dans laquelle la décision d_i consiste à admettre que $X \sim P^{X(j)}$. Le problème revient alors à choisir une règle de décision (mixte, dans le cas général) $m \in \Delta_M$ qui maximise la probabilité d'une classification correcte.

En **théorie bayésienne**, si Π est une **loi a priori** définie à partir d'une **tribu de parties** de Θ, le problème équivaut à minimiser le **risque de T. BAYES** :

(15) $R_{\Pi}(m) = 1 - \Sigma_{i=1}^{k} \Pi_i E_{\theta(i)} m(d_i / X)$.

Le **théorème de HOEL-PETERSON** en donne la solution optimale.

(vii) D'un point de vue terminologique, le terme de **classification** peut recevoir des sens différents. Ainsi :

(a) dans certains cas, il est nécessaire d'admettre que l'unité (individu) ω appartient à un (k+1)-ème ensemble (population) tout-à-fait nouveau (voire inconnu). Il arrive ainsi que l'on mette à jour des « découvertes », ie que des unités ω ne puissent pas se rattacher aux classes Ω_1 ,..., Ω_k déjà connues, et qu'elles définissent donc une nouvelle classe Ω_{k+1} (ce que l'on associe généralement à un **effet de suprise**) ;

(b) on distingue parfois (eg P. DAGNELIE) entre :

$(b)_1$ **« problème de classement »**, ou **problème d'affectation**, ou encore **problème d'imputation**, consistant, comme ci-dessus, à affecter (ou à imputer) à une classe Ω_i d'une partition (donnée) Π_Ω de Ω un élément ω (ou, plus généralement, un ensemble d'éléments A : eg un échantillon) de Ω, ie à admettre que ω (resp A) est extrait d'une classe donnée Ω_i de Ω.

Cette situation correspond, notamment, au problème de **classification « statistique »**, ici simplement appelé problème de classification.

$(b)_2$ **« problème de classification »**, consistant à définir des classes dans un ensemble Ω d'éléments (eg unités statistiques), généralement à l'aide de « grandeurs » observées sur ces éléments. Le but est de regrouper (ou de répartir) les éléments en classes homogènes, et hétérogènes entre elles (cf **hétérogénéité**, **homogénéité**, **partition**).

Cette seconde situation correspond souvent à des problèmes de **classification automatique**.

Dans ces deux cadres, la **théorie des parties floues** peut conduire à définir des **partitions** floues : ie certains éléments de Ω peuvent appartenir à deux classes ou davantage.