CLASSIFICATION (I7, K9)

(06 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Le terme de **classification** correspond aussi bien :

- (a) à la **procédure statistique** (action) consistant à établir une classification ;
- (b) à la classification (résultat) ainsi obtenue en fin de procédure.

On distingue généralement entre :

(a) les méthodes de **classification statistique**, ou les méthodes de **discrimination**, et les méthodes de **classification automatique** ou de « **classification formelle** » ;

(b) les méthodes de classification de type hiérarchique (cf **hiérarchie**, **analyse hiérarchique**) et celles de type non hiérarchique. Ces dernières sont, en général, réalisées par partitionnements successifs : **méthode des nuées dynamiques**, méthodes agrégatives (cf **agrégation**, **agrégation selon la variance**).

Toutes ces méthodes utilisent des notions de **distance** ou de **proximité**, et autres notions analogues : **écart**, **norme**, **dissimilarité**, **association**, **corrélation**. La définition des classes d'une classification utilise des notions d'**homogénéité** de différents types.

(i) Soit Ω un **ensemble d'unités statistiques** (**population**, univers cosmique, fabrication industrielle, etc) et $\Pi_{\Omega} = (\Omega_m)_{m=1,\dots,M}$ une **partition** finie de Ω en classes Ω_m (catégories sociales, galaxies, chaînes de production, etc).

Un **problème de classification** consiste à déterminer une partition en classes Ω_m homogènes. On considère généralement une liste de **variables** (certaines étant des **variables qualitatives**, d'autres des **variables numériques**), qui sont **observables** sur ω et généralement représentées par un vecteur (**mesures**) $x = \xi(\omega) \in \mathbf{R}^K$, pour les K variables numériques, et par une suite $y = \eta(\omega) \in \mathcal{Y}^G$ pour les G variables non numériques. Ces variables sont souvent considérées comme des **variables aléatoires**.

La méthode de classification retenue consiste en l'**observation** sur chaque élément ω des caractères ξ et η , les résultats de cette observation étant resp notés $x = \xi(\omega)$ et $y = \eta(\omega)$.

(ii) Dans l'approche « statistique » des classifications, on doit distinguer entre :

(a) un **problème de classification** ;

(b) un **problème de discrimination**, aussi appelé problème d'**analyse discriminante**, **problème de classement** ou encore (plus rarement) **problème d'identification**. Un problème de discrimination consiste à imputer un élément donné $\omega \in \Omega$ (individu, planète, pièce, etc) à une classe Ω_m d'une partition prédéfinie $\Pi_{\Omega} = (\Omega_m)_{m=1,\dots,M}$ de Ω . L'appartenance de ω à une classe Ω_m n'est pas connue a

priori et l'on cherche à la déterminer (ie à « discriminer » entre les Ω_m) au vu des variables observées sur l'élément ω .

De façon équivalente, dans un **problème d'affectation** les Ω_m ($m \in N_M^*$) sont autant d'ensembles distincts et ω représente un élément extrait au hasard de l'un d'entre eux, Ω_m , l'indice m étant inconnu (cf aussi **mélange de lois**).

La partition est donc une inconnue dans le premier cas (classification), une donnée dans le second (discrimination).

Il arrive souvent qu'un problème de discrimination succède à un problème de classification. Ainsi (sociologie : eg paléontologie), des « catégories » étant établies, on cherche à adjoindre à l'une d'entre elles une nouvelle unité observée : cette unité peut soit être affectée à une catégorie, soit nécessiter la création d'une nouvelle catégorie (effet de surprise).

(iii) Le problème central d'une classification consiste à classer les unités en groupes aussi « distincts » que possible, ie :

(a) partitionner la population en classes homogènes (cf **homogénéité**), aussi hétérogènes que possible entre elles. Ce partitionnement peut s'effectuer de façon stricte ou de façon « floue », selon l'objectif ou la tolérance ;

(b) définir des groupes homogènes d'unités dont l'ensemble reconstitue la population donnée ;

(c) ou encore définir des sous-populations homogènes, ce qui est différent de la **dissection d'un mélange de lois**.

Deux approches principales visent à résoudre un problème de classification.

(iv) La **première approche**, ou **classification automatique**, consiste à définir, à partir des observations, un ensemble de partitions successives de Ω . Celles-ci sont :

(a) soit de plus en plus fines (depuis la **partition grossière** $\{\Omega, \emptyset\}$ jusqu'à la **partition discrète**, ou **partition triviale**, $\{\omega\}_{\omega \in \Omega}$;

(b) soit de moins en moins fines : depuis la partition discrète jusqu'à la partition grossière (cf **classification emboîtée**).

On parle aussi de **classification hiérarchique** (cf **analyse hiérarchique**) lorsque les méthodes précédentes comportent des étapes successives (descendantes dans le premier cas, ascendantes dans le second) (cf aussi **hiérarchie**), et de **classification non hiérarchique** sinon :

(a) une méthode descendante peut eg séparer progressivement les unités les plus dissemblables entre elles (cf **dissimilarité**) : méthode dite par **partitionnement** ou par **segmentation** ;

(b) une méthode ascendante peut eg regrouper progressivement les unités les plus proches entre elles : méthode dite par **recollement** ou par **agglutination**.

Ces méthodes utilisent, à titre de **variables de classification**, soit des variables (ξ, η) du type précédent, soit des **distances** (ou pseudo-distances, ou quasi-distances, etc), soit des **indicateurs de différences** à caractère discriminant (cf **indice de dissimilarité, indice de similarité**) ;

(v) La **seconde approche**, ou **classification statistique**, adopte, de façon explicite, le cadre de la **théorie de la décision (décision statistique)**.

Dans le cas de **variables numériques** ξ , si $(\Omega_m, \mathcal{F}_m, P_m)$ désigne, $\forall m \in N_M^*$, l'**espace probabilisé** associé à la population Ω_m , on cherche à définir une partition $(B_m)_{m=1, \dots, M}$ de l'**espace d'observation** $\mathbf{R}^K$ tq :

(1) si $x = \xi(\omega) \in B_m$, on prend la **décision** $\omega_m \in \Omega_m$.

On suppose donnée une **matrice de coût** $C = (c_m)_{(l,m)} \in M_M(\mathbf{R})$ dans laquelle c_{lm} représente le coût dû à une décision erronée $\omega \in \Omega_l$ alors que $\omega \in \Omega_m$ (avec, naturellement, $c_{mm} = 0, \forall m$). La probabilité de décider $\omega \in \Omega_l$ alors que $\omega \in \Omega_m$ est :

$$(2) \quad p_{lm} = \int \mathbf{1}(B_l) dP_m^\xi(x),$$

(où, $\forall m \in N_M^*$, P_m^ξ est l'image de P_m par ξ) et le coût moyen des décisions erronées portant sur Ω_m s'explique selon :

$$(3) \quad \bar{c}_m = \sum_{l=1}^M p_{lm} c_{lm}.$$

Par suite :

(a) la **théorie bayésienne** suppose données des **probabilités a priori** π_m selon lesquelles ω provient de Ω_m (avec $m \in N_M^*$, $\pi_m \geq 0$ et $\sum_{l=1}^M \pi_m = 1$). Le coût moyen d'ensemble s'écrit :

$$(4) \quad \bar{c} = \sum_{m=1}^M \pi_m \bar{c}_m.$$

Ce coût dépend de la partition $(B_m)_{m=1, \dots, M}$ et il s'agit de déterminer une partition (solution du problème) qui rende $\bar{c}$ minimum.

On suppose que les P_m^ξ sont dominées par une même **mesure positive** σ -finie μ (cf **famille dominée de lois**) et l'on pose $dP_m^\xi(x) / d\mu(x) = f_m(x)$. La **probabilité conditionnelle** (a posteriori) que $\omega \in \Omega_l$ sachant que $[\xi = x]$ est alors donnée par le **théorème de BAYES**, d'où :

$$(5) \quad p_l(x) = \pi_l \cdot f_l(x) / \sum_{m=1}^M \pi_m f_m(x).$$

Le coût moyen (a posteriori) correspondant s'écrit alors :

$$(6) \quad \gamma_l(x) = \sum_{m=1}^M p_m(x) c_{lm}.$$

On doit donc déterminer l'indice $l \in N_M^*$ tq $\gamma_l(x)$ soit minimum. Les régions B_m ($m \in N_M^*$) sont finalement déterminées par les conditions :

$$(7) \quad B_m = \{x \in \mathbf{R}^K : \gamma_m(x) < \gamma_l(x), \forall l = 1, \dots, m-1, m+1, \dots, M\}.$$

En l'absence d'information, les probabilités a priori sont parfois choisies uniformes ($p_m = 1 / M, \forall m$) (cf **loi uniforme discrète, loi a priori non informative**) ;

(b) la **méthode « minimax »** doit utiliser un autre principe que le théorème de BAYES lorsque les probabilités a priori π_m ne sont pas connues. Elle procède à partir des coûts moyens $\bar{c}_m$. Ainsi, lorsqu'on peut trouver une suite $(\pi_m)_{m=1, \dots, M}$ tq $\bar{c}_m = \bar{c}_0$ (**constante**), $\forall m \in N_M^*$, la **solution minimax** du problème de classification est donnée par (cf **règle minimax**) :

$$(8) \quad B_m = \{x \in \mathbf{R}^K : \pi_l \cdot f_l(x) < \pi_m \cdot f_m(x), \forall l \neq m\},$$

avec encore : $f_m(x) = dP_m^\xi(x) / d\mu(x)$.