

ANALYSE DISCRIMINANTE (K7)

(03 / 08 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **problème de discrimination** est un **problème d'imputation** : l'objectif est de rattacher un élément donné à un **ensemble** d'éléments donné a priori figurant parmi plusieurs tels ensembles. Il se rattache notamment au **problème de reconnaissance des formes**.

(i) L'**analyse discriminante** est une branche de la **Statistique** qui consiste :

(a) « **analyse exogène** » : soit à déterminer l'**ensemble** A_i parmi plusieurs ensembles $(A_i)_{i=1,\dots,n}$ donnés, auquel appartient un élément a donné (souvent appelé « individu »). Cet élément est de même « nature », ou de même « type », que les éléments des ensembles considérés ;

(b) « **analyse endogène** » : soit, de façon comparable, à imputer un élément a d'un ensemble A à l'une des classes A_i d'une **partition** $\Pi_E = (A_i)_{i=1,\dots,n}$, donnée a priori, de cet ensemble A (cf **classification**, **reconnaissance des formes**). Cet élément est aussi de même « nature », ou de même « type », que les éléments des classes considérées ;

(c) « **analyse externe** » : soit à imputer un élément donné a, « extérieur » au **contexte statistique** (ou à l'**information** disponible : **données** diverses) :

(c)₁ ou bien à un **ensemble** A_i parmi plusieurs ensembles $(A_i)_{i=1,\dots,n}$ « concurrents » donnés (« analyse comparative ») ;

(c)₂ ou bien à l'une des classes A_i d'une **partition** $\Pi_E = (A_i)_{i=1,\dots,n}$, donnée a priori, d'un ensemble donné A (cf **classification**, **reconnaissance des formes**).

Dans la situation (c), l'élément considéré peut être de même « nature », ou de même « type », que les autres éléments pris en compte. Mais il peut aussi être de nature, ou de type, différent (« **effet de surprise** » ou « **effet de découverte** », parfois « **effet d'aubaine** »).

(ii) Les ensembles A, ou **parties ensemblistes** A_i , considéré(e)s peuvent être quelconques, ou être des **échantillons** ou encore des **populations**. Ils (elles) sont ainsi constitué(e)s d'**unités élémentaires** assimilables à des **unités statistiques** : eg **unités expérimentales**, **unités de sondage**.

Les classes de la partition Π_E sont des sous-échantillons (disjoints entre eux) ou des sous-**populations** (disjointes entre elles), et l'on cherche à imputer une **unité** donnée à l'une des ces classes (cf aussi **classement**, **problème de classement**).

Dans le cas de **parties floues**, l'imputation d'un élément peut concerner deux ensembles flous, voire davantage, et chaque imputation se voit attribuer un **degré de vraisemblance**, ou **degré d'appartenance**.

(iii) Pour effectuer l'analyse, on se réfère généralement à une liste $\zeta = (\zeta_1, \dots, \zeta_J)$ constituée de J **variables observables** sur chaque unité élémentaire a, d'où chaque observation individuelle $\zeta(a) = (\zeta_1(a), \dots, \zeta_J(a)) = (z_{1,a}, \dots, z_{J,a})$.

Une analyse discriminante peut donc être univariée ($J = 1$) ou multivariée ($J > 1$). Elle peut concerner plusieurs éléments (les individus) à classer.

(iv) Le contexte précédent tient compte de l'impossibilité éventuelle d'observer directement des **populations**, ou les classes d'une partition relative à une population : en effet, on peut généralement « prélever » dans chacune d'elle un (des) **échantillon(s)**.

Par suite, le **problème de discrimination**, ou **problème de classement**, revient à affecter une observation (supplémentaire) à un (et un seul) des ensembles considérés (resp à une (et une seule) des classes considérées). Ceci s'effectue généralement en minimisant la **probabilité de classement erroné** et la décision est prise à l'aide d'une **fonction discriminante**, ou **statistique discriminante** : eg la **fonction discriminante de FISHER**.

(v) Cette analyse est généralement basée sur un **principe bayésien** : le **statisticien** adopte une **probabilité a priori** (éventuellement uniforme) de classement, et la probabilité a posteriori permet de calculer les **fréquences** de classement erroné précédentes (cf **loi a posteriori**).

(vi) En **analyse des données**, l'**analyse discriminante** porte sur l'étude de la **dépendance** entre une **variable qualitative** et des **variables quantitatives**. Cette approche, qui peut être considérée comme un cas particulier de l'**analyse en composantes principales** ou de l'**analyse canonique**, consiste principalement à rechercher un **codage optimal** de la variable qualitative considérée.

(vii) Selon de **domaine de connaissance**, des exemples d'imputation peuvent être les suivants :

(a) physique : corps céleste $\leftrightarrow$ galaxie, en fonction de leur **structure** (composition), de leur masse ou de leur dynamique, etc ;

(b) biologie (généalogie) : descendant $\leftrightarrow$ ascendant, en fonction de leurs génômes ;

(c) écologie : type d'organisation animale (territoire, dominance) $\leftrightarrow$ types dominants dans d'autres zones, en fonction de la structure d'implantation, des déplacements, etc ;

(d) psychologie : catégorie de comportement d'un groupe $\leftrightarrow$ typologie de comportements grégaires (coopération vs conflit, partage vs exclusivité, etc) ;

(e) sociologie (paléontologie) : groupe d'ossements $\leftrightarrow$ espèce particulière (relativement à une période de temps).