Chapitre 20

Analyse discriminante

L'**analyse discriminante** est une technique statistique ayant pour objectif de décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire, *etc*.) d'un ensemble d'observations (individus, exemples, *etc*.) à partir d'une série des variables prédictives (descripteurs, variables exogènes, *etc*.). La construction d'outils discriminants répond à des **objectifs opérationnels**.

Le champ d'application de l'analyse discriminante est large :

— la médecine ;

— le marketing ;

— l'analyse financière des entreprises ;

— l'agronomie ;

— la reconnaissance de formes ;

— la reconnaissance de la parole ;

— l'archéologie ;

— la météorologie ;

— *etc*.

La discrimination est un terme neutre utilisé pour distinguer des groupes (des ensembles) ayant des caractéristiques propres suffisamment fortes pour constituer des groupes.

La perception humaine est souvent plus efficace que les techniques mathématiques pour définir des groupes. Qu'en est-il si le nombre de variables se multiplie ? Le développement informatique permet à l'analyse discriminante de traiter les cas complexes.

L'analyse discriminante débute par la **description des oppositions entre les groupes *a priori*** à partir de variables caractérisant les individus, appelées **descripteurs**.

À partir des variables, on peut construire un **indicateur synthétique** dont les valeurs permettent d'opposer les groupes. Cela permet d'établir une **règle d'affectation** à un des groupes pour tout nouvel individu à examiner et dont l'appartenance à un des groupes est inconnue.

L'analyse discriminante vise à construire un **outil d'aide à la décision** qui permet d'assigner à une des catégories toute observation nouvelle caractérisée par ses descripteurs, mais dont on ignore le groupe d'appartenance.

L'analyse discriminante utilise la **méthode expérimentale** dans la mesure où elle confronte constamment la réalité et les résultats du modèle mis en place. Elle fournit la possibilité d'évaluer la performance de la règle d'assignation de l'objet, par le calcul de **taux de bons classements**. La règle d'assignation à un groupe, fournie par l'outil discriminant, s'accompagne généralement d'un certain pourcentage d'erreurs. Cela permet d'estimer une probabilité *a posteriori* d'appartenir à un groupe une fois sa caractérisation connue par l'outil discriminant. De fait, la discrimination est une **classification supervisée**.

Soit une population d'individus E. Chaque individu peut être placé dans différentes situations appelées **état de la nature**. Ces états sont **exclusifs** les uns des autres. Ils permettent de définir une **partition** de la population en nombre fini k de groupes : $E_1, \ldots, E_k$. L'ensemble des états de la nature est noté : $\Theta) \{\theta_1, \theta_2, \ldots, \theta_k\}$. Il existe k états de la nature correspondant aux k groupes définis. Chaque individu est caractérisé par des variables appelés descripteurs. Elles peuvent être de différentes natures : qualitatives, quantitatives ou mixtes.

Les analyses discriminantes possibles sont très nombreuses :

— l'analyse discriminante de E. M. Fisher [Fisher, 1936] ;

— la discrimination logistique (avec une régression logistique) ;

— l'analyse discrimante décisionnelle (ou l'analyse discrimante prédictive), qui effectue une prévision avec une technique de *scoring*. Il s'agit de construire une **fonction de classement** (ou des règles d'affectation) qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables prédictives. Elle entre dans un **cadre probabiliste**, comme la **technique de la régression logistique** [1], fondée sur un modèle de régression binomiale et inventée par Joseph Berkson en 1944 et en 1951. Elle constitue un cas particulier de modèle linéaire généralisé. Elle se base sur un choix binaire (0 et 1).

Joseph Berkson (1899-1982)

$$\ln\left(\frac{\Pr(X\,|1)}{\Pr(X\,|0)}\right) = a_0 + a_1 x_1 + \ldots + a_j x_j \tag{20.1}$$

ou

$$\ln\left(\frac{\Pr(1\,|X)}{1-\Pr(1\,|X)}\right) = b_0 + b_1 x_1 + \ldots + b_j x_j \tag{20.2}$$

1. ou modèle logit

$$\Pr(1|X) = \frac{e^{b_0+b_1x_1+\ldots+b_jx_j}}{1+e^{b_0+b_1x_1+\ldots+b_jx_j}} \tag{20.3}$$

— l'analyse discriminante descriptive (ou analyse factorielle discriminante) ;

— la méthode DISQUAL inventé par Gilbert Saporta [Saporta, 1975] ;

Gilbert Saporta (né en 1946)

— les réseaux de neurones ;

— la méthode d'Emmanuel Parzen [Parzen, 1962] ;

Emmanuel Parzen (1929-2016)

— la méthode des k plus proches voisins ;

— les arbres de discrimination.

L'analyse discriminante est utilisée pour :

— de la statistique exploratoire [2] ;

— de l'analyse de données [3] ;

— de la reconnaissance de formes [4] ;

Exemple. La reconnaissance optique des caractères [5] (R.O.C.) ou l'océrisation

— de l'apprentissage automatique [6] ;

— de la fouille de données [7] ;

— *etc*.

Les traitements de variables se distinguent :

— pour les valeurs manquantes (ou censurées) ;

— pour les valeurs aberrantes ;

— pour les valeurs extrêmes.

2. En anglais : *Exploratory Data Analysis*
3. En anglais : *Data Analysis*
4. En anglais : *Pattern Recognition*
5. En anglais : *Optical Character Recognition* (O.C.R.)
6. En anglais : *Machine Learning*
7. En anglais : *Data Mining*

Bibliographie

[Fisher, 1936] FISHER, E. M. (1936). Linear discriminant analysis. Statistics & Discrete Methods of Data Sciences, (392):1–5.

[Parzen, 1962] PARZEN, E. (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33(3):1066–1075.

[Saporta, 1975] SAPORTA, G. (1975). Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, Paris.