L'analyse discriminante

M. Calciu

Introduction

Présentation

L'analyse discriminante connue dans la pratique marketing sous le nom "scoring" essaye de déterminer la contribution des variables qui expliquent l'appartenance des individus à des groupes. Deux ou plusieurs groupes sont comparés, pour déterminer s'ils différent et pour comprendre la nature de ces différences. En marketing il est important d'identifier des caractéristiques qui déterminent l'appartenance à des groupes, qui permettraient de distinguer entre:

- utilisateurs permanents et occasionnels d'un produit;

- acheteurs d'une marque et les acheteurs de marques concurrentes;

- clients loyaux et aloyaux

- vendeurs bons, médiocres et mauvais [1]

Exemple:

Par exemple suite à un concours de vente on a réussi de séparer trois groupes de vendeurs. Pour identifier les éléments (les variables) qui permettent de distinguer entre un bon et un mauvais vendeur, on a appliqué un questionnaire aux participants a ce concours dans le quel on s'est intéressé d'une manière prioritaire à quatre caractéristiques : le nombre de contacts avec des nouveaux clients ; la proportion de contacts avec rendez-vous d'avance; les coups de téléphone aux prospects et le nombre de nouveau comptes visités [2] .

Tableau 1 Actions des vendeurs concernant les nouveaux clients

Modèle géométrique

Cas à deux variables

Pour simplifier, on prend deux groupes de sujets (vendeurs) mesurés sur deux variables X et Y. Une première variable indiquant le nombre de contacts avec des nouveaux clients sera notée Y et la deuxième variable, proportion des contacts par rendez-vous sera notée par X.

Les réponses des individus à ces questions, sont représentées dans la figure 1. Les deux questions représentent les variables selon lesquelles on aimerait arriver à distinguer les deux groupes.

Figure 1

En analysant le graphique on se rend compte que sur chacune des axes représentés par les deux variables il y a une région importante d'incertitude, dans la quelle pour les même valeurs des variables on trouve des individus appartenant aux deux groupes (bons et mauvais vendeurs).

Objectif

Le but de l'analyse discriminante serait de trouver un nouvel axe comme combinaison linéaire des variables qui permettrait de réduire cette zone d'incertitude. Un tel axe dans l'exemple donné est illustré dans la figure 2.

Figure 2

Cas général

Imaginons maintenant, d'une manière plus générale, que ces variables sont centrées pour l'ensemble des deux groupes.

On illustre dans un espace à deux dimensions les coordonnées de chaque sujet sur les variables en question et on trace les lignes contours de 95 % (1,96) Ou 68 % (1 a). Les hypothèses de normalité des variables conjointes et d'homogénéité des variances et covariances sont maintenues [3] .

La figure 3 en est un exemple.

Fig.3 Lignes contours de deux échantillons de variables conjointes homogènes et normalement distribuées.

Dans la démarche habituelle de l'analyse, on éprouvera d'abord le besoin de déterminer s'il faut considérer les moyennes ou centroïdes de chaque population comme distinctes: L'analyse de variance multiple est un instrument

bien indiqué pour cette tâche. Il apparaît ensuite intéressant, connaissant les scores d'un éventuel sujet sur les deux variables, de déterminer, à l'intérieur d'un pourcentage d'erreur, l'appartenance de ce sujet à l'une ou l'autre de ces populations.

Zones d'incertitude et attribution d'appartenance

Si l'on se base, pour une telle attribution d'appartenance, sur la connaissance du seul score X, on voit que l'incertitude de la décision va de a à b et touche les sujets situés dans la région hachurée de la figure 4.

Fig.4 Région d'incertitude basée sur la variable X

Fig. 5 Région d'incertitude basée sur la variable Y

La région sure

Si cette décision est prise en tenant compte des deux variables, on constate que la région sure correspond alors à la surface hachurée

Fig. 6 Région d'incertitude basée sur deux variables

Région minimum d'incertitude

Cette seconde façon de procéder, quoique meilleure que la première, peut comporter encore de grosses imprécisions (sauf cas particuliers) car la véritable région d'incertitude est celle de l'intersection des deux ellipses

Fig. 7 Région minimum d'incertitude

Recherche des axes discriminants

On propose de considérer une nouvelle variable qui soit une combinaison linéaire des précédentes; géométriquement, cette nouvelle variable est représentée par un axe sur lequel on projette les divers points des groupes de sujets. Aux fins d'illustration, on limite le nombre de variables de départ ainsi que le nombre de groupes à deux; dans la pratique, le modèle mathématique dont il sera question à l'article suivant ne comporte pas de telles limites.

L'axe est appelé axe de la fonction discriminante. Les points projetés sur cet axe se distribuent normalement pour chacun des groupes; les valeurs de cette fonction comprises entre les lignes pointillés correspondent à la région d'incertitude. On désire que l'axe occupe une position telle que la projection des

points donne lieu au minimum de superposition des divers groupes de sujets.

La figure 8 illustre la situation particulière de la recherche de la fonction de deux variables qui discrimine au maximum deux groupes de cas.

Maximiser le pouvoir discriminant

La qualité de la discrimination est liée à la superposition des deux distributions de projections sur l'axe. On peut mesurer la qualité de la dispersion à la grandeur du rapport de la variance entre les moyennes à la variance à l'intérieur d'un groupe variance inter-groupe / variance intra-groupe.

Ce rapport est analogue au F de l'analyse de variance. On suppose que la variance des scores à l'intérieur de chaque groupe répond au critère d'homogénéité de telle sorte que cette variance intra est la moyenne des variances intra des groupes considérés.

Un rapport maximum est lié non seulement à la grandeur de son numérateur mais aussi à l'étroitesse du dénominateur: la variance inter atteindra son maximum pour l'axe parallèle au segment joignant les centroides tandis que la variance intra sera minimum pour un axe perpendiculaire à l'axe principal des ellipses; c'est en une position intermédiaire que se situe le rapport maximum des variances inter et intra.

Figure 8. Modèle géométrique de la fonction discriminante.

Resultats

Les coefficients de la fonction discriminante

Utilisation des coefficients

Une manière de déterminer quelles sont les variables qui discriminent entre les deux types de gagnants aux concours de ventes est de construire un index, basé sur les valeurs des caractéristiques mesurées, qui sépare les deux groupes, formant une combinaison linéaire de ces dernières du genre:

ou sont des coefficients arbitraires. En analyse discriminante les coefficients sont dérivés de telle manière que la variation des scores de Y entre les (deux) groupes soit si large que possible et la variation des scores de Y à l'intérieur des groupes (within group ou intra-groupe) soie si petite que possible.

Autrement dit on calcule les coefficients qui maximisent le rapport variance inter-groupe / variance intra-groupe.

Ce ci fait les groupes aussi distincts que possible du point de vue des nouveaux scores de l'index.

Pour l'exemple analysé les coefficients discriminants sont et la combinaison linéaire qui différencie de manière maximale entre les groupes est [4]

Y = 0,059 X1 + 0,063 X2 + 0,034 X3 - 0,032 X4,

Calcul des scores sur les axes discriminants

Ayant les coefficients discriminants on peut calculer le score de chaque vendeur, si ce score est plus proche de la moyenne des scores du groupe des gagnants du grand prix, l'individu sera affecté à ce groupe si non il était affecté à l'autre groupe.

On peut observer que l'approche de l'analyse discriminante est proche de celui de la régression. Dans les chacun des cas on essaye d'expliquer (prévoir) une variable dépendante par une combinaison linéaire de variables indépendantes. En régression la variable expliquée est continue. En analyse discriminante la variable dépendante est l'appartenance à un groupe. On peut transformer le problème d'analyse discriminante pour de groupes en problème de régression en utilisant une variable muette (dummy) comme variable dépendante. Les coefficients de régression résultants seront proportionnels à ceux obtenus par l'analyse discriminante.

Scores calculés des gagnants du grand prix et du prix de consolation utilisant la fonction discriminante

Y = 0,059 X1 + 0,063 X2 + 0,034 X3 - 0,032 X4,

Tableau 2 - Calcul de scores discriminants

Interprétation de la fonction discriminante

Tester la différentiation obtenue

Dans une approche rigoureuse avant d'interpréter la fonction discriminante, on doit tester si au niveau des scores discriminants on obtient une différentiation significative entre les groupes. Cela se fait en appliquant un test en F aux valeurs de la statistique D2 de Mahalanobis (qui mesure la distance de chaque

case à la moyenne du groupe, tout en permettant des axes corrélés et des unités de mesure différentes).

Les coefficients discriminants

Classification des individus utilisant la fonction discriminante

Calcul de cutting score

Pour classifier les individus dans un des groupes on doit fixer un score (cutting score) qui joue le rôle de frontière entre les groupes. Normalement c'est la moyenne des scores des deux groupes. Si les groupes sont de dimensions égales le cutting score (YCS) est égale à la moyenne des moyennes des scores des groupes.

Si les groupes ne sont pas égaux on utilisera une moyenne pondérée du genre:

où 1 et 2 sont les scores discriminants moyens et n1 et n2 sont les dimensions des groupes (on observe que la moyenne de chaque groupe est pondérée avec la dimension de l'autre.

Attribution aux groupes

Dans le cas du concours des vendeurs, une règle de décision simple serait de classifier un vendeur dans le groupe des gagnants du grand prix si son score est plus proche de la moyenne des scores du groupe des gagnants du grand prix, que de la moyenne des scores du groupe des gagnants du prix de consolation, si non il sera affecté au groupe des gagnants du prix de consolation.

Tableau 3 - L'appartenance de groupe prédite utilisant la règle de classification simple

La qualité de la classification

La conformité de cette classification prédictive avec la réalité est illustrée par le tableau suivant, appelé matrice des confusions

Tableau 4 - Matrice des confusions

En générale cette matrice est un tableau de contingence g x g (où g est le

nombre de groupes), en ligne figurent les appartenances réelles et en colonnes les affectations par le modèle. On peut y repérer le nombre d'affectations correctes et erronées [5] . Le hit rate est le pourcentage d'affectations correctes par rapport au nombre total d'individus. Pour que le modèle présente un intérêt, il faut que le hit score soit suffisamment élevé.

Dans le cas de deux groupes à effectifs égaux [6] , une procédure de répartition purement aléatoire entraînerait 50 % d'affectations correctes. La différence entre le hit score et 50 % mesure ainsi la qualité du modèle. Le caractère significatif de cette différence est repéré à l'aide de l'expression:

Z = (p 0,5)/[(0,5)(0,5)/n]1/2

où n est le nombre d'individus. Si Z est supérieur à 1,64, le modèle a significativement mieux réussi à classer les individus qu'un processus aléatoire, à un seuil de 95 % [7] .

Critères du maximum chance et proportional chance

Quand les groupes sont de dimensions différentes le hit rate ne peut plus être comparé au critère du 50% dans ce cas on peut utiliser deux critères: le critère de la probabilité maximum (maximum chance) et le critère de la probabilité proportionnelle (proportional chance). Le critère maximum chance considère que tout individu choisit aléatoirement doit être classé comme appartenant au plus grand groupe. Le critère proportional chance est donné par la somme des carrés des proportions de chaque groupe par rapport au nombre totale d'individus (dans le cas de deux groupes Cpro = p2 + (1 - p)2).

Développement mathématique du modèle [9]

Matrice de covariances intra-groupes

Soit une matrice X de scores centrés sur v variables. Celle-ci est partitionnée en g sous-matrices Di de ni cas.

À chaque partition correspond un centroïde de mi = (mi1,...miv), une matrice de dispersion (sommes des carrés des écarts et des produits des écarts aux moyennes) Wi ainsi qu'une matrice de covariances Vi = Wi /(ni - 1).

Les matrices de covariances de chaque groupe étant homogènes, il y a intérêt à considérer la matrice de dispersion au sein des groupes W = W1 + · · · + Wg à laquelle correspond la matrice moyenne de covariances intra groupes V= W/(ni-g).

L'ensemble des centroïdes mj constitue la matrice

Soit = 0.

Matrice de covariances inter-groupes

La matrice correspondante des covariances est dite d'intergroupes et s'obtient ainsi:

B = M'M/(g - 1)

La fonction discriminante y recherchée est celle qu'on obtient au moyen d'une combinaison linéaire k de la matrice X de telle sorte qu'à une variance intra groupes donnée corresponde un maximum de la variance inter groupes. On peut représenter ainsi la dispersion des cas et moyennes des trois groupes sur l'axe y de la fonction discriminante.

Figure 9 - Dispersion de trois groupes sur l'axe de la fonction discriminante.

La fonction discriminante expression générale

La fonction discriminante ayant pour expression générale y = Xk, les projections sur y des centroïdes, c'est-à-dire des moyennes des groupes, seront Mk et la variance de ces moyennes ou variance inter sera k'M'Mk/(g - 1) = k'Bk. De la même façon, la variance intra groupes sera k'Vk.

On désire considérer simultanément une dispersion maximum des moyennes pour une dispersion donnée des cas au sein des groupes. On choisit d'arrêter la condition d'une variance intra unité, c'est-à-dire k'Vk = 1, qui s'ajoute au multiplicateur indéterminé de Lagrange pour former l'équation

F = k'Bk - (k'Vk - 1)

Critères de maximisation

On calcule alors les valeurs du vecteur k qui maximisent F:

= 2Bk - 2 Vk = 0

= Bk - Vk = 0

= (B - V)k = 0

On sait que celle équation comporte une solution différente de zéro pour |B - V| = 0.

Ces équations, solutions du problème de la fonction discriminante présentent donc une grande ressemblance avec celles du problème de l'analyse factorielle. On leur donne la forme habituelle en pré multipliant par l'inverse de V:

$(V^{-1}B - I)k = 0$

et

$|V^{-1}B - I| = 0$

C'est le problème de la recherche des vecteurs latents k d'une matrice carrée non symétrique V-1B. Il y a intérêt à faire en sorte de revenir a une matrice symétrique. On a vu (théorème 7 ...) qu'une matrice non symétrique M peut être écrite comme la somme d'une matrice symétrique S et d'une matrice non symétrique A : M = S + A. Par multiplications et additions appropriées de lignes, on rend diagonale la matrice symétrique et on remplace V-1B par $DV^{-1/2}BDv^{-1/2}$ qui a la propriété d'être symétrique et d'avoir les mêmes racines et vecteurs latents que V-1B.

L'expression de la solution de la fonction discriminante devient donc:

$(DV^{-1/2}BDv^{-1/2} - I)k = 0$

$| DV^{-1/2}BDv^{-1/2} - I | = 0$

La résolution de l'équation

$| DV^{-1/2}BDv^{-1/2} - I | = 0$

comporte un nombre de solutions égal au minimum suivant: soit le nombre de groupes moins un, soit le nombre de variables. Chacune de ces valeurs est introduite dans l'équation

$(DV^{-1/2}BDv^{-1/2} - iI)ki = 0$

entraînant un nombre correspondant de solutions ki. On a vu antérieurement que les vecteurs ki sont orthogonaux, c'est-à-dire indépendants: la matrice K des vecteurs ki répond donc à la relation $K'K = I$ lorsque normée aux racines latentes.

Le calcul se fait à partir de la valeur des matrices B et V des covariances inter et intra. Si on procède à partir des matrices de dispersion, la solution serait la même sauf pour les racines latentes qui s'en trouveraient multipliées par le rapport

$(g - 1) / (ni - g)$.

Plusieurs auteurs présentent la fonction discriminante comme étant définie par le vecteur k tel que le rapport de la variance inter à la variance intra soit maximum:

Rotation des axes des fonctions discriminantes

Interprétation

Une fonction discriminante est définie par un vecteur colonnes de coefficients appliqués, par une combinaison linéaire, aux variables étudiées; ces coefficients sont dits bruts ou standard (beta weights) selon qu'ils s'appliquent à des variables brutes ou standard. Le but de telles combinaisons linéaires est de séparer au maximum les groupes les uns des autres; en plus d'être indépendantes les unes des autres, les fonctions discriminantes ont un pouvoir de discrimination qui décroît d'une fonction à l'autre. Le nombre de fonctions discriminantes est le plus petit des deux possibilités suivantes: nombre de variables ou nombre de groupes moins un.

Méthodes de rotation

On a vu qu'à beaucoup de points de vue, les fonctions discriminantes présentent une grande analogie avec les facteurs mis en évidence par l'une ou l'autre technique de l'analyse factorielle. En particulier, on peut souhaiter identifier par un nom chacune de celles-ci, en se basant sur les contributions des variables, telles qu'exprimées de façon comparable par les coefficients standard. Comme en analyse factorielle, une telle identification n'est pas toujours facile, à moins de procéder à une rotation des axes des fonctions tout en maintenant constantes les positions relatives des cas et des moyennes ou centroïdes. Une rotation VARIMAX est proposée en option dans le programme SPSS; on obtient ainsi des coefficients qui sont le plus possible voisins de 1 pour les uns, et de zéro pour les autres. L'avantage qu'on en tire est celui d'une facilité plus grande d'interprétation, mais cependant on y perd la connaissance de l'ordre des fonctions quant à leur pouvoir de discrimination. C'est pourquoi il est suggéré de n'utiliser qu'avec prudence la rotation des axes des fonctions discriminantes (voir Klecka, dans SPSS, p. 444-445).

Méthodes de sélection des variables

Plusieurs méthodes peuvent être utilisées dans le choix des variables à inclure dans l'édification des fonctions discriminantes. Celle dont il a été question jusqu'à maintenant, et qui consiste à considérer toutes les variables à la fois, est dite méthode directe.

On peut aussi faire appel à une approche hiérarchique (stepwise) où les variables sont introduites une à une selon leur capacité décroissante à mettre en évidence la différence entre les groupes. Au cours des sélections successives, il est possible que des variables déjà entrées perdent leur pouvoir de discrimination: la raison en est une redondance d'information, c'est-à-dire que le pouvoir de discrimination de cette variable est désormais inclus dans quelque combinaison de nouvelles variables retenues.

Donc à chaque étape de l'analyse, on procède à l'élimination des variables devenues inutiles. Dans le programme SPSS ce test de variable apparaît sous le titre F-TO-REMOVE.

Divers critères, mettant l'accent sur l'un ou l'autre aspect de la dispersion des groupes, sont utilisés pour la sélection de variables:

a) le test de Wilks vise à minimiser un rapport où entrent en considération la dispersion des centroïdes et la cohésion des cas au sein des groupes: il est semblable à un test multivarié F sur les différences entre les centroïdes;

b) plusieurs tests, reliés à la notation de distance de Mahalanobis, visent à maximiser l'écart entre les deux groupes les plus rapprochés (les méthodes MAHAL, MAXMINF, MINRESID du programme SPSS sont des variantes de cette approche);

c) la méthode de Rao consiste à choisir la variable qui contribue le plus à une distance généralisée, évaluée sur les variables précédentes.

Pour tous ces critères, une variable est sélectionnée lorsque son rapport F partiel dépasse une valeur critique, c'est-à-dire lorsque sa contribution à la dispersion additionnelle des centroïdes est statistiquement significative: dans le programme SPSS ce F est dit F-TO-ENTER

Test de signification

On peut poursuivre, jusqu'à exhaustion, l'extraction des fonctions discriminantes. Mais comme dans le cas des composantes principales, l'intérêt des fonctions additionnelles va décroissant. Dans beaucoup d'applications on ne dépasse pas deux ou trois fonctions afin de tirer parti de la facilité et de l'intérêt d'une illustration de la position des groupes de sujets dans un espace à trois dimensions et moins.

L'effet de discrimination de la fonction i par rapport à toutes les fonctions est exprimé par la proportion (Hope, p. 117-120; Cooley et Lohnes, p.248-250)

Ce rapport exprime la proportion de la variance expliquée par chaque fonction discriminante. Cependant cette proportion ne conduit pas à une décision

statistique au sens habituel du terme. On recourt souvent à un autre indicateur.

On montre que:

,où p= nombre de fonctions discriminantes,

peut être utilisé pour exprimer la capacité de discrimination d'un ensemble

de variables (ce paramètre est similaire à de l'article 11.3.0. de Laforge). De même pour les fonctions au-delà de la k-ième fonction:

Ce lambda ( ) est donc une mesure de l'inverse de la puissance discriminative expliquée par les fonctions discriminantes à venir. La signification de la discrimination des fonctions restantes k à p, à la suite de l'acceptation des k premières, peut se calculer au moyen de l'approximation de Bartlett:

avec d.l. = (v-k) (g-k-1)

où v: nombre de variables

g: nombre de groupes

et

Si pour ces fonctions discriminantes (k + 1) à p, on obtient une valeur de x2 qui ne dépasse pas le seuil critique, on considère que les k premières fonctions calculées suffisent seules à expliquer de façon significative les écarts entre les groupes.

Remarques et résumé

Discussion

L'analyse discriminante peut être vue comme un cas spécial d'analyse factorielle. Mais le but diffère: il s'agit de faire ressortir au maximum les différences entre des groupes mesurés dans un espace multidimensionnel, en projetant chaque cas dans l'espace unidimensionnel d'un petit nombre de fonctions linéaires orthogonales.

Cette opération fait suite habituellement à celle de l'analyse de variance multivariée où, en présence d'une situation où plusieurs groupes sont mesurés sur plusieurs variables, on s'intéresse d'abord à déterminer s'il y a différence significative entre les groupes. Dans le cas de résultats positifs, il devient intéressant de déterminer, parmi les variables, celles qui sont responsables

dans un ordre décroissant d'importance des différences entre les groupes: c'est le but de l'analyse discriminante. Une exploitation plus poussée des résultats conduit à leur utilisation dans le but de classifier (en se donnant comme objectif une probabilité minimum d'erreurs) des nouveaux sujets dans les divers groupes: l'étude de cette technique fait l'objet du chapitre suivant.

Le rôle de l'analyse discriminante peut être envisagé de deux façons quant à l'attribution des qualificatifs d'indépendance et de dépendance, aux variables mesurées sur les populations visées et aux fonctions discriminantes. En sciences d'exploration, en général, les populations sont considérées comme variables indépendantes (predictor) et les fonctions discriminantes comme variables dépendantes (critères). En sciences expérimentales, ces rôles se trouvent renversés.

L'analyse discriminante consiste donc à projeter dans un sous-espace approprie des échantillons de mesures multidimensionnelles. L'interprétation de cette opération peut être faite en termes (voir Cooley et Lohnes, p. 243) soit du nombre et de l'importance relative des fonctions discriminantes retenues, soit de la localisation dans l'espace discriminant des populations étudiées.