Analyse discriminante

L'analyse discriminante vise à classer des individus dans différents groupes à partir de variables explicatives continues. C'est une méthode supervisée : les groupes sont connus dans les données d'apprentissage et l'objectif est d'apprendre une règle de classification optimale.

Notation

Soit $X = (X_{ij}) \in \mathbb{R}^{n \times p}$ un matrice de données, où n est le nombre d'individus dans l'échantillon, p est le nombre de variables et X_{ij} est la valeur de la je variable pour le ie individus.

On suppose qu'il y a K groupes dans la population et que chaque individu appartient à l'un des K groupes. Pour $k \in \{1, \dots, K\}$, on note I_k l'ensemble des individus du groupe k, n_k le nombre d'observations dans I_k. On a donc $\sum_{k=1}^{K} n_k = n$.

Objectif

Nos observations sont dans $\mathbb{R}^p$. Pour faire de la classification à partir de $X_1, \dots, X_p$, on doit partionner $\mathbb{R}^p$ en K sous-ensembles de sorte que chacun des K sous-ensembles soit associé à l'un des K groupes. Dans le cas de l'analyse discriminante, l'idée est de passer de $\mathbb{R}^p$ à $\mathbb{R}$ en calculant, pour chaque observation, un score $f(X_1, \dots, X_p) \in \mathbb{R}$ et ensuite utiliser ce score pour déterminer le groupe d'appartenance (et donc de partionner $\mathbb{R}$). Le score proposé par Fisher est une combinaison linéaire des variables, i.e.

$$f(X_1, \dots, X_p) = a^\top X + b = a_1 X_1 + \dots + a_p X_p + b.$$

Cette fonction de score nous permet de regrouper les individus d'un même groupe dans des zones proches du score et de séparer autant que possible les différents groupes selon ce score. On en déduira K intervalles de décision $S_1, \dots, S_K$ associés aux groupes.

Remarque

Sans perte de généralité, on peut choisir

$$-b = a_1\overline{X}_1 + \dots + a_p\overline{X}_p = a^\top \overline{X}$$

ce qui permet de centrer les variables en enlevant le vecteur de moyenne

$$\overline{X} = \left(\overline{X}_1, \dots, \overline{X}_p\right).$$

Il ne reste plus qu'à choisir le vecteur $a = (a_1, \dots, a_p)$.

Critère de Fisher

L'idée centrale de Fisher est d'optimiser le rapport entre la variabilité inter-groupes et la variabilité intra-groupe en fonction du vectuer a. Dit autrement, on voudrait choisir le vecteur a de sorte que les scores soient, à la fois, très différents entre les groupes et très similaires à l'intérieur d'un groupe. On s'intéresse donc à la variabilité des scores à l'intérieur des groupes et entre les groupes.

Notons

— S, la matrice de variance-covariance totale donnée par

$$S = \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^\top \quad \text{où} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

— W, la matrice de variance-covariance intra-groupe donnée par

$$W = \sum_{k=1}^{K} \sum_{i \in I_k} (X_i - \overline{X}_k)(X_i - \overline{X}_k)^\top \quad \text{où} \quad \overline{X}_k = \frac{1}{n_k}\sum_{i \in I_k} X_i;$$

— B, la matrice de variance-covariance inter-groupe donnée par

$$B = \sum_{k=1}^{K} n_k (\overline{X}_k - \overline{X})(\overline{X}_k - \overline{X})^\top.$$

Étant donné que $a \in \mathbb{R}^p$, on a :

$$\mathrm{Var}(f(X_1, \dots, X_p)) = \mathrm{Var}(a^\top X) = a^\top \mathrm{Var}(X) a.$$

Comme S est un estimateur de la variance totale de X, nous pouvons estimer cette variance par

$$\widehat{\mathrm{Var}}(f(X_1, \dots, X_p)) = \frac{1}{n} a^\top S a.$$

La base de l'analyse discriminante repose sur le fait que

$$S = W + B.$$

Preuve que $S = W + B$

On peut prouver ce résultat en considérant la définition des matrices S, W et B. La moyenne de la variable j pour tous les individus de l'échantillon est

$$\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}.$$

La moyenne de la variable j pour les individus du groupe k est

$$\overline{X}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} X_{ij}.$$

La somme des carrés totale est

$$s_{jj'} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ij'} - \overline{X}_{j'}).$$

On tirerait de la matrice S une estimation de $\mathrm{Cov}(X_j, X_{j'})$ si toutes les observations provenaient d'un même groupe. On définit $s_{jj'}$ comme étant

$$s_{jj'} = w_{jj'} + b_{jj'},$$

où

$$w_{jj'} = \sum_{k=1}^{q} \sum_{i \in I_k} (X_{ij} - \overline{X}_{kj})(X_{ij'} - \overline{X}_{kj'}),$$

$$b_{jj'} = \sum_{k=1}^{q} n_k (\overline{X}_{kj} - \overline{X}_j)(\overline{X}_{kj'} - \overline{X}_{j'}).$$

Preuve :

1. Poser $X_{ij} - \overline{X}_j = X_{ij} - \overline{X}_{kj} + \overline{X}_{kj} - \overline{X}_j$, dans la définition de $s_{jj'}$, iden pour $X_{ij'}$.
2. Développer les produits.
3. Remplacer $\sum_{i=1}^{n}$ par $\sum_{k=1}^{q} \sum_{i \in I_k}$.
4. Faire les simplications appropriées.

Ainsi, en remplaçant S par $W+B$, on obtient

$$\widehat{\mathrm{Var}}(a^\top X) = \frac{1}{n} a^\top S a = \frac{1}{n}\left(a^\top W a + a^\top B a\right).$$

Le critère de Fisher s'écrit donc, pour $a \in \mathbb{R}^p$,

$$J(a) = \frac{a^\top B a}{a^\top W a}.$$

Celui-ci peut se réécrire de façon équivalente comme

$$J(a) = \frac{a^\top B a}{a^\top S a}.$$

On cherche à maximiser $J(a)$. Ce problème peut être reformulé des façons suivantes, toutes équivalentes :

1. Maximiser $J(a)$ sous la contrainte que $a^\top a = 1$.
2. Maximiser $a^\top B a$ sous la contrainte que $a^\top S a = 1$.
3. Maximiser $c^\top S^{-1/2} B S^{-1/2} c$ sous la contrainte que $c^\top c = 1$, où $c = S^{1/2} a$.

En réécrivant la troisième formulation, on obtient

$$c^\top \left(S^{-1/2} B S^{-1/2}\right) c \quad \text{s.c.} \quad c^\top c = 1,$$

on peut prendre $a = S^{-1/2} c$, où c est un vecteur propre normé associé à λ_1, la première valeur propre de $S^{-1/2} B S^{-1/2}$. De façon équivalente, de la deuxième formulation, on peut prendre a, un vecteur propre normé associé à λ_1 la première valeur propre de $S^{-1}B$. Notons que comme, pour λ une valeur propre quelconque de $S^{-1/2} B S^{-1/2}$,

$$S^{-1/2} B S^{-1/2} c = \lambda c \quad \text{et} \quad a = S^{-1/2} c,$$

alors

$$S^{-1/2} B a = \lambda S^{1/2} a \Rightarrow S^{-1} B a = \lambda a.$$

Les valeurs propres de $S^{-1}B$ et de $S^{-1/2} B S^{-1/2}$ sont donc les mêmes et on peut facilement passer des vecteurs propres de $S^{-1}B$ à ceux de $S^{-1/2} B S^{-1/2}$.

Fonction discriminante

La fonction discriminante de Fisher est donc

$$f(x) = a^\top (x - \overline{X}),$$

où a est le vecteur propre normé associé à la plus grande valeur propre de $S^{-1}B$. Les scores, i.e. la représentation des observations dans $\mathbb{R}$, sont données par $U_i = a^\top (X_i - \overline{X})$. Ces scores sont des combinaisons linéaires des variables et ils maximisent le rapport variance intra-groupe / variance inter-groupes.

Remarque

On peut aussi prendre $U_i = a^\top X_i$, car ajouter la même constante à toutes les observations $i = 1, \dots, n$ ne fait que décaler la représentation dans $\mathbb{R}$.

Pouvoir discriminant

Puisque la matrice $S^{-1/2}BS^{-1/2}$ est symétrique et définie positive, ses valeurs propres sont toutes réelles et positives. De plus, on a que $S^{-1}Ba = \lambda_1 a$. Ainsi,

$$Ba = \lambda_1 Sa \Rightarrow a^\top Ba = \lambda_1 a^\top Sa \Rightarrow \lambda_1 = \frac{a^\top Ba}{a^\top Sa}.$$

On a donc $0 \leq \lambda_1 \leq 1$. La valeur propre λ_1 peut donc être vue comme le pouvoir discriminant de f :

— $\lambda_1 = 1 \Rightarrow a^\top Ba = a^\top Sa$, donc 100% de la variabilité est entre les groupes et il n'y a aucune variabilité à l'intérieur des groupes. Cela représente le cas idéal.

— $\lambda_1 = 0 \Rightarrow a^\top Ba = 0$, donc il n'y a aucune variabilité entre les gorupes et toute la variabilité est concentrée à l'intérieur des groupes. Cela représente le cas où l'analyse discriminante est inutile.

Règle de classification

Après avoir estimer la fonction discriminante $f(x)$, on peut calculer le score moyen de chaque groupe k défini comme étant

$$m_k = a^\top \left(\overline{X}_{k1}, \dots, \overline{X}_{kp}\right)^\top,$$

où $\overline{X}_{kj}, j = 1, \dots, p$ est la moyenne de la je variable pour les individus appartenant au ke groupe.

Considérons maintenant une nouvelle observation $X_0 \in \mathbb{R}^p$ qui n'était pas dans le jeu de données lors de l'estimation de $f(x)$. Pour classer ce nouvel individu dans un groupe de la population, on calcule son score $f(X_0) = a^\top X_0$. Ensuite, on l'assigne au groupe $k^\star$ dont le score moyen m_k est le plus proche de $f(X_0)$, i.e. le groupe tel que

$$\left|a^\top X_0 - m_{k_0}\right| = \min_{1 \leq k \leq K} \left|a^\top X_0 - m_k\right|.$$

Autrement dit, on a

$$k^\star = \arg\min_k \left|a^\top X_0 - m_k\right|.$$

En applicant cette règle à l'échantillon $X_1, \dots, X_n$, on peut estimer les risques de mauvaise classification avec la matrice de confusion.

Cas particulier de la classification binaire

Dans le cas où $K = 2$, il est possible de calculer explicitement la valeur de vecteur propre a. Soit

$$C = \sqrt{\frac{n_1 n_2}{n}}(\overline{X}_1 - \widetilde{X}_2),$$

$\overline{X}_i, i = 1, 2$ sont les vecteurs moyens des caractériques dans chaque groupe, alors

$$B = CC^\top, \quad a = S^{-1}C.$$

Supposons que

$$m_1 = a^\top \overline{X}_1 > a^\top \overline{X}_2 = m_2.$$

Alors, on classe un individu X_0 dans le premier groupe si

$$a^\top X_0 > \overline{m} = \frac{m_1 + m_2}{2} = a^\top \left(\frac{\overline{X}_1 + \overline{X}_2}{2} \right),$$

et dans le second groupe sinon. Cette règle de classification est équivalente à

$$(\overline{X}_1 - \overline{X}_2)^\top S^{-1} X_0 > \frac{1}{2}(\overline{X}_1 - \overline{X}_2)^\top S^{-1}(\overline{X}_1 + \overline{X}_2).$$