Data Mining 2 – Modélisation Statistique et Apprentissage
Master Statistique & Econométrie
UT1 Sciences Sociales, Toulouse
E-thème 2 **17 mars 2018**
Enseignant: Gilles Cohen

Analyse discriminante

1 Analyse discriminante

L'analyse discriminante [1] est utilisée dans des situations où l'on connait à priori les classes. Le but de l'analyse discriminante est de classer une observation, ou plusieurs observations, dans ces groupes connus. Par exemple, dans le crédit scoring, une banque sait de par son expérience passée qu'il y a des bons clients (qui remboursent leur prêt sans aucun problème) et des mauvais clients(qui ont eu des difficultés à rembourser leurs prêts). Lorsqu'un nouveau client demande un prêt, la banque doit décider si oui ou non accorder le prêt. L'information dont dispose la banque se trouve dans deux ensembles de données : des observations multivariées sur les deux catégories de clients (comprenant âge, salaire, status marital, le montant du prêt, ...). La règle de discrimination doit alors classer le client dans l'un des deux groupes (ou classes) existants, et l'analyse discriminante devrait évaluer le risque de mauvaise classement. On va présenter la fonction de discrimination du maximum de vraisemblance (MV) et le discriminant linéaire de Fisher.

Formellement, le problème est d'essayer d'affecter une observation à l'une des populations $\Pi_j, j = 1, 2, \ldots, J$. Une règle discriminante correspond à une séparation de l'espace d'entrée (généralement $\mathbb{R}^p$) en ensembles disjoints R_j telle que si une nouvelle observation se trouve dans la région R_j , elle est identifiée comme un membre de la population Π_j. La qualité de la règle de discrimination peut être jugée sur la base de l'erreur de mauvaise classement. Si les fonctions de densité des populations Π_j sont connues, on peut alors

1. Dans l'analyse discriminante ou classement on affecte les objets à des groupes préétablis ; l'analyse discriminante fixe ainsi des règles pour déterminer la classe des objets. La classification est le travail préliminaire au classement, à savoir la recherche des classes "naturelles" dans le domaine étudié. On trouvera souvent dans la littérature classifieur à la place de "'classeur" qui est une mauvaise traduction du mot classifier en anglais. Les sciences biologiques utilisent également le terme de "taxinomie" pour désigner l'art de la classification et les sciences de l'homme le terme de "typologie".

facilement en dériver une règle de discrimination basée sur l'approche du maximum de vraisemblance.

1.1 Règle discriminante du maximum de vraisemblance

Chaque population $\Pi_j, j = 1, \ldots, J$ peut être décrite par une fonction de densité $f_j(x)$. La règle du maximum de vraisemblance affecte la nouvelle observation x à la population Π_k, maximisant la vraisemblance $L_k(x) = f_k(x) = max_{i=1,\ldots,J} f_i(x)$. Formellement, les ensembles $R_j, j = 1, \ldots, J$ donnés par la règle discriminante du MV sont :

$$R_j = \{x : f_j(x) \geq f_i(x) \text{ pour } i = 1, \ldots, J\}.$$

En pratique, les ensembles R_j sont construites à partir des densités inconnues. Si les densités ont une forme connue, i.e., une distribution normale, il suffit alors d'estimer les paramètres inconnus.
Si $\Pi_j = \mathcal{N}_p(\mu_j, \Sigma), j = 1, 2..$ On fait l'hypothèse que la matrice de variance Σ est définie positive. La vraisemblance de l'observation x dans chaque population $\Pi_j, j = 1, 2$ est

$$L_j \quad = \quad f_i(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma^{-1}(x-\mu_j)\right\}$$

Selon la règle du maximum de vraisemblance, on affecte x à la population Π_j avec la plus grande vraisemblance. En omettant la constante $|2\pi\Sigma|^{-1/2}$ et en prenant les logarithmes, le problème de maximisation peut être résolu de façon équivalente en minimisant :

$$\begin{aligned} \delta^2(x, \mu_j) &= (x-\mu_j)^T\Sigma^{-1}(x-\mu_j) \\ &= \{\Sigma^{-1/2}(x-\mu_j)\}^T\Sigma^{-1/2}(x-\mu_j). \end{aligned}$$

$\delta^2(x, \mu_j)$ est le carré de la distance de Mahalanobis entre x et μ_j. Donc, dans le cas d'une distribution normale avec même matrice de covariance, la règle du MV affecte x au groupe le plus proche au sens de la distance de Mahalanobis.
Pour $J = 2$, l'observation x est affectée à Π_1 si

$$(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \leq (x-\mu_2)^T\Sigma^{-1}(x-\mu_2).$$

En réarrangeant les termes on obtient :

$$\begin{aligned} 0 &\geq -2\mu_1^T\Sigma^{-1}x + 2\mu_2^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 \\ 0 &\geq 2(\mu_2-\mu_1)^T\Sigma^{-1}x + (\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2) \\ 0 &\leq (\mu_1-\mu_2)^T\Sigma^{-1}\{x - \frac{1}{2}(\mu_1+\mu_2)\} \\ 0 &\leq \alpha^T(x-\mu) \end{aligned}$$

où $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ et $\mu = \frac{1}{2}(\mu_1 + \mu_2)$. Il s'ensuit que dans le cas de deux populations multinormales, la règle de discrimination peut s'écrire comme :

$$R_1 = \{x : \alpha^T(x-\mu) \geq 0\}$$

1.2 Règle discriminante de Bayes

La qualité de la règle discriminante du MV peut être améliorée si l'on a une connaissance à priori sur la probabilité des populations. Soit π_j la probabilité à priori de la classe j. On a $\sum_{j=1}^{J}\pi_j = 1$. La règle discriminante de Bayes affecte x à la population Π_k qui donne la plus grande valeur de $\pi_i f_i(x), \pi_k f_k(x) = \max_{i=1,\ldots,J}\pi_i f_i(x)$. La règle discriminante de Bayes peut être formellement définie par :

$$R_j = \{x : \pi_j f_j(x) \geq \pi_i f_i(x) \text{ pour } i = 1, \ldots, J\}.$$

La règle de Bayes est identique à la règle discriminante du MV si $\pi_j = 1/J$

1.3 Discriminant linéaire LDA

Pour deux populations univariée normalement distribuée $\Pi_1 = \mathcal{N}(\mu_1, \sigma)$ et $\Pi_2 = \mathcal{N}(\mu_2, \sigma)$, la règle du MV peut s'écrire comme

$$\begin{aligned} R_1 &= \left\{x : (\mu_1-\mu_2)\left(x - \frac{\mu_1+\mu_2}{2}\right) \geq 0\right\} \\ R_1 &= \left\{x : \text{sign}(\mu_1-\mu_2)\left(x - \frac{\mu_1+\mu_2}{2}\right) \geq 0\right\} \\ R_1 &= \left\{x : \text{sign}(\mu_1-\mu_2)x \geq \text{sign}(\mu_1-\mu_2)\frac{\mu_1+\mu_2}{2}\right\} \end{aligned}$$

En faisant l'hypothèse que $\mu_1 < \mu_2$, on obtient

$$R_1 = \left\{x : x \leq \frac{\mu_1+\mu_2}{2}\right\}$$

i.e., on classe x comme appartenant à R_1 si il est plus proche de μ_1 que de μ_2.

1.4 Discriminant quadratique QDA

On suppose que les deux populations normales ont des variances différentes, $\Pi_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ et $\Pi_2 = \mathcal{N}(\mu_2, \sigma_2^2)$, on affecte x à R_1 si $L_1(x) > L_2(x)$ où la vraisemblance est :

$$L_i(x) \quad = \quad (2\pi\sigma_i^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)\right\}$$

$L_1(x) \geq L_2(x)$ est équivalent à $L_1(x)/L_2(x) \geq 1$ et on obtient

$$\begin{aligned}
\frac{\sigma_2}{\sigma_1} \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\right\} &\geq 1 \\
\log\frac{\sigma_2}{\sigma_1} - \frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right] &\geq 0 \\
\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right] &\leq \log\frac{\sigma_2}{\sigma_1} \\
x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) &\leq 2\log\frac{\sigma_2}{\sigma_1}.
\end{aligned}$$

Si $\sigma_1 = \sigma_2$, la plupart des termes dans la formule ci-dessus disparait et l'on retrouve le résultat précédent.

1.5 Discriminant linéaire de Fisher

La règle du discriminant linéaire de Fisher est basé sur la maximisation du rapport entre la variance intra-groupe et inter-groupe de la projection $a^T x$. Si

$$\mathcal{Y} \quad = \quad \mathcal{X}a$$

représente une combinaison linéaire des observations, alors la somme totale des carrés de y, $\sum_{i=1}^n (y_i - \bar{y})^2$, est égale à

$$\mathcal{Y}^T\mathcal{H}\mathcal{Y} \quad = \quad a^T\mathcal{X}^T\mathcal{H}\mathcal{X}a = a^T\mathcal{T}a \tag{1}$$

avec la matrice de centrage $\mathcal{H} = \mathcal{I} - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$ et $\mathcal{T} = \mathcal{X}^T\mathcal{H}\mathcal{X}$. La matrice de centrage est une matrice symétrique et idempotente, qui lorsque multipliée par un vecteur a le même effet que de sosutraire la moyenne des composantes

du vecteurs de chaque composantes.
Les deux matrices importantes sont la matrice identité $\mathcal{I}$ $(n \times n)$:

$$\mathcal{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

et la matrice de centrage $\mathcal{H}$ $(n \times n)$:

$$\mathcal{H} = \mathcal{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix}$$

On suppose que l'on a des échantillons $\mathcal{X}_j, j = 1, \dots, J$ provenant de J populations. Soit $\mathcal{Y} = \mathcal{X}a$ et $\mathcal{Y}_j = \mathcal{X}_j$ a une combinaison linéaire des observations. La somme des carrés intra-groupe est donnée par :

$$\sum_{j=1}^{J} \mathcal{Y}_j^T \mathcal{H}_j \mathcal{Y}_j \quad = \quad \sum_{j=1}^{J} a^T \mathcal{X}_j^T \mathcal{H}_j \mathcal{X}_j a = a^T \mathcal{W} a,$$

où $\mathcal{H}_j$ représente la matrice de centrage $n_j \times n_j$. La somme des carrés inter-groupes est :

$$\sum_{j=1}^{J} n_j (\bar{y}_j - \bar{y})^2 \quad = \quad \sum_{j=1}^{J} n_j \{a^T(\bar{x}_j - \bar{x})\}^2 = a^T \mathcal{B} a,$$

où $\bar{y}_j$ et $\bar{x}_j$ représente la moyenne de $\mathcal{Y}_j$ et $\mathcal{X}_j$ et $\bar{y}$ et $\bar{x}$ la moyenne de $\mathcal{Y}$ et $\mathcal{X}$.

La somme totale des carrés (1) est la somme des carrés inter-groupes et de la somme des carrés entre groupes, i.e.,

$$a^T \mathcal{T} a \quad = \quad a^T \mathcal{W} a + a^T \mathcal{B} a$$

L'idée de Fisher est de choisir un vecteur de projection a qui maximize le rapport :

$$\frac{a^T \mathcal{B} a}{a^T \mathcal{W} a}$$

Le vecteur a qui maximise $a^T\mathcal{B}a/a^T\mathcal{W}a$ est le vecteur propre de $\mathcal{W}^{-1}\mathcal{B}$ qui correspond à la plus grande valeur propre.
Finalemment, l'observation x est classée dans le groupe j, qui est le plus proche de la projection $a^T x$,

$$R_j \quad = \quad \{x : |a^T(x - \bar{x}_j) \leq |a^T(x - \bar{x}_i)| \text{ pour } i = 1, \ldots, J\}$$

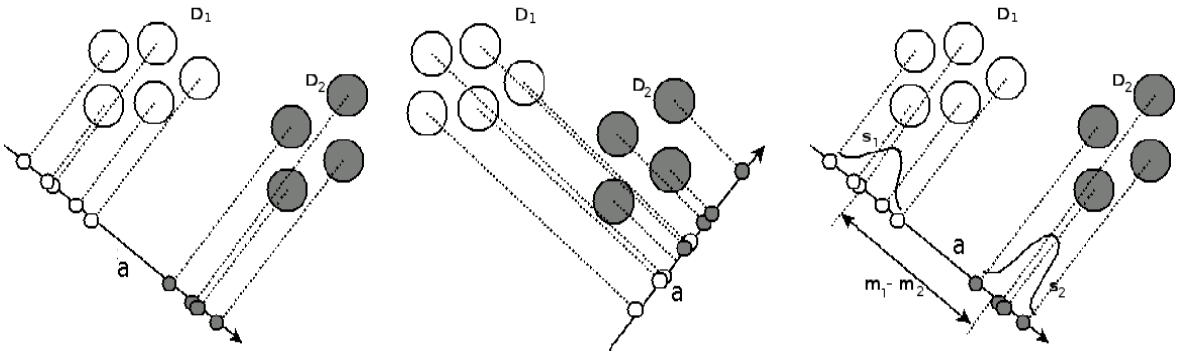

FIGURE 1 – A gauche – Projection $\mathbb{R}^2 \to \mathbb{R}$ sur la direction donnée par le vecteur **a**. On peut trouver un seuil pour le classement. Au centre – Projection $\mathbb{R}^2 \to \mathbb{R}$ sur la direction donnée par le vecteur **a**. La projection n'est pas bonne pour faire de la classement. A droite – Le mieux est de séparer les classes – les distances entre les projections de m_1 et de m_2 doivent être maximisées et celles entres les projections des variances doivent être minimisées.

1.6 Estimation de paramètres

En pratique on a besoin d'estimer π, μ_k et Σ_k pour spécifier complètement la distribution jointe $P(\mathbf{x}, y)$.

- $\hat{\pi}_k = \hat{P}(y = k) = \frac{N_k}{N}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$
- $\hat{\Sigma}_k = \frac{1}{n_k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$

Dans le cas où la matrice de covariance est commune, on appliquera l'estimation du maximum de vraisemblance.

Cas où J=2

La fonction de vraisemblance est :

$$\begin{aligned}
\log P(D|M) &= \sum_{i=1}^{N} \log p(\mathbf{x}_i, y_i) \\
&= \sum_{i=1}^{N} \log\{[\pi.\mathcal{N}(\mathbf{x}_i|\mu_i, \Sigma)]^{y_i}[(1-\pi).\mathcal{N}(\mathbf{x}_i|\mu_0, \Sigma)]^{1-y_i}\} \\
&= \sum_{i=1}^{N} \{y_i \log \pi + y_i \log \mathcal{N}(\mathbf{x}_i|\mu_i, \Sigma) + (1-y_i)\log(1-\pi) + (1-y_i)\log \mathcal{N}(\mathbf{x}_i|\mu_0, \Sigma)\}
\end{aligned}$$

où $\mathcal{N}(\mathbf{x}_i|\mu_i, \Sigma)$ tirage des x_i selon loi normale de paramètres μ_i, Σ. On commence par estimer π. On considère pour cela la partie qui contient π, on a :

$$\sum_{i=1}^{N} \{y_i \log \pi + (1-y_i)\log(1-\pi)\}$$

On prend la dérivée par rapport à π :

$$\frac{1}{\pi}\sum_{i=1}^{N} y_i - \frac{1}{1-\pi}\sum_{i=1}^{N}(1-y_i)$$

puis on l'égale à 0 et l'on pose $N_1 = \sum_{i=1}^{N} y_i$ et $N_2 = \sum_{i=1}^{N}(1-y_i)$, on a :

$$\begin{aligned}
\frac{N_1}{\pi} &= \frac{N_2}{1-\pi} \\
\pi &= \frac{N_1}{N_1+N_2}
\end{aligned}$$

On estime maintenant μ_1 (l'estimation de μ_0 est identique), On considère la partie qui contient μ_1, on a alors ;

$$\begin{aligned}
\sum_{i=1}^{N} \log \mathcal{N}(\mathbf{x}_i|\mu_1, \Sigma) &= \sum_{i=1}^{N} y_i \frac{-(\mathbf{x}_i-\mu_1)^T\Sigma^{-1}(\mathbf{x}_i-\mu_1)}{2} + \text{ const} \\
&= \sum_{y_i=1} \frac{-(\mathbf{x}_i-\mu_1)^T\Sigma^{-1}(\mathbf{x}_i-\mu_1)}{2} + \text{ const}
\end{aligned}$$

On prend la dérivée par rapport à μ_1 et on l'égale à 0, on a :

$$\begin{aligned}
\sum_{y_i=1} \Sigma^{-1}\mathbf{x}_i - \mu_1 &= 0 \\
\mu_1 &= \frac{1}{N_1}\sum_{y_i=1} \mathbf{x}_i
\end{aligned}$$

On a de façon similaire :

$$\mu_1 = \frac{1}{N_2} \sum_{y_i=0} \mathbf{x}_i$$

Pour finir, on va maintenant estimer σ. On rappelle que $Tr(A) = \sum_i A_{ii}$ est la trace de la matrice A (somme des éléments de la diagonale). Elle satisfait la propriété de la permutation cyclique :

$$tr(ABC) = tr(CAB) = tr(BCA)$$

On peut alors en déduire l'astuce de la trace, qui réordonne le produit scalaire intérieur $x^T Ax$ comme suit

$$x^T Ax = tr(x^T Ax) = tr(xx^T A)$$

On prend la partie qui contient Σ, on a alors :

$$-\frac{N_1}{2} \ln|\Sigma| - \frac{1}{2} y_i (\mathbf{x}_i - \mu_1)^T \Sigma^{-1} (\mathbf{x}_i - \mu_1) - \frac{N_2}{2} \ln|\Sigma| - \frac{1}{2}(1 - y_i)(\mathbf{x}_i - \mu_0)^T \Sigma^{-1} (\mathbf{x}_i - \mu_0)$$

$$\begin{aligned}
&= -\frac{N}{2} \ln|\Sigma| - \frac{N_1}{2} Tr(\Sigma^{-1} S_1) - \frac{N_2}{2} Tr(\Sigma^{-1} S_2) \\
&= -\frac{N}{2} \ln|\Sigma| - \frac{N}{2} Tr\left(\Sigma^{-1} \left(\frac{N_1}{N} S_1 + \frac{N_2}{N} S_2\right)\right) \\
&= -\frac{N}{2} \ln|\Sigma| - \frac{N}{2} Tr(\Sigma^{-1} S)
\end{aligned}$$

où S_i est la matrice de dispersion (scatter matrix) pour chaque classe

$$S_i = \sum_{\boldsymbol{x} \in \{1,2\}}^{n} (\boldsymbol{x} - \boldsymbol{m}_i)\,(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

On prend les dérivées de cette expression par rapport à Σ. On utilise les résultats suivants :

$$\begin{aligned}
\frac{\partial}{\partial A} Tr(BA) &= B^T \\
\frac{\partial}{\partial A} \log|A| &= A^{-T}
\end{aligned}$$

et on égale à 0, on a alors :

$$\Sigma = S = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$$

1.7 Exemples

— Règle discriminante de Bayes

1. $\pi_1 = \pi_2 = 0.5$
2. $\mu_1 = (0,0)^T$, $\mu_2 = (2,-2)^T$.
3. $\Sigma = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 0.5625 \end{pmatrix}$
4. Frontière de décision :
   $5.56 - 2.00x_1 + 3.56x_2 = 0.0$

— Règle discriminante du MV (Iris de Fisher).

1. $N_1 = N_2 = 50$
2. $\mu_1 = (5, 3.4)^T$, $\mu_2 = (6, 2.8)^T$.
3. $S_1 = \begin{pmatrix} 0.12 & 0.10 \\ 0.10 & 0.14 \end{pmatrix}$ $S_2 = \begin{pmatrix} 0.26 & 0.08 \\ 0.08 & 0.10 \end{pmatrix}$
   d'où $S_p = \frac{50S_1+50S_2}{98}$
   $= \begin{pmatrix} 0.19 & 0.09 \\ 0.09 & 0.12 \end{pmatrix}$
4. Donc $d = S_p^{-1}(\mu_1 - \mu_2)$
   $= \begin{pmatrix} 0.19 & 0.09 \\ 0.09 & 0.12 \end{pmatrix}^{-1} \begin{pmatrix} -1.0 \\ 0.6 \end{pmatrix} = \begin{pmatrix} -11.4 \\ 14.1 \end{pmatrix}$
   $\mu = \frac{1}{2}(\mu_1 - \mu_2) = \begin{pmatrix} 5.5 \\ 3.1 \end{pmatrix}$
5. Frontière de décision : Affecter x à Π_1 si

$$\begin{aligned} -11.4(x_1 - 5.5) + 14.1(x_2 - 3.1) &> 0 \\ -11.4x_1 + 14.1x_2 + 19.0 &> 0 \end{aligned}$$