

Analyses discriminantes quadratique et linéaire

T. Dencœux

1 Introduction

On suppose dans ce chapitre que le vecteur de caractéristique \mathbf{X} suit, conditionnellement à chaque classe ω_k ($k = 1, \dots, c$), une loi normale multidimensionnelle d'espérance $\boldsymbol{\mu}_k$ et de variance Σ_k . En faisant différentes hypothèses sur les paramètres de ces lois (notamment sur les matrices de variance), on obtient différentes expressions de la règle de Bayes, d'où l'on déduit différentes règles de décision en remplaçant les paramètres théoriques par leurs estimations.

2 Analyse discriminante quadratique

2.1 Modèle

Considérons tout d'abord le cas général où la distribution de \mathbf{x} dans chaque classe est caractérisée par des paramètres μ_k et Σ_k différents. On alors

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}(\det \Sigma_k)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

Pour simplifier, nous nous placerons dans ce chapitre dans le cas des coûts $\{0, 1\}$. En notant a_k l'action de choisir la classe ω_k , la règle de Bayes s'écrit alors $g^*(\mathbf{x}) = a_{k^*}$ avec

$$\begin{aligned} k^* &= \arg \max_k \mathbb{P}(\omega_k | \mathbf{x}) \\ &= \arg \max_k \pi_k f_k(\mathbf{x}) \\ &= \arg \max_k g_k(\mathbf{x}), \end{aligned}$$

avec

$$\begin{aligned} g_k(\mathbf{x}) &= \ln f_k(\mathbf{x}) + \ln \pi_k & (1) \\ &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \ln(\det \Sigma_k) + \ln \pi_k - \frac{p}{2} \ln(2\pi). & (2) \end{aligned}$$

Les fonctions $g_k(\mathbf{x})$ qui servent à définir la règle de décision sont appelées *fonctions discriminantes*. Ici, ce sont des formes quadratiques : on parle de *fonctions discriminantes quadratiques*. Les régions de décision sont séparées par des frontières d'équations $g_k(\mathbf{x}) = g_\ell(\mathbf{x})$. En dimension quelconque, ces variétés sont des quadratiques (hypersphères, hyperellipsoïdes, hyperparaboloïde, etc.). En dimension 2, ce sont des coniques (cercles, ellipses, paraboles, hyperboles, droites).

2.2 Estimation des paramètres

En pratique, les paramètres π_k , $\boldsymbol{\mu}_k$ et Σ_k du modèles sont inconnus, mais on dispose d'un ensemble d'apprentissage $\mathcal{L} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, supposé être une réalisation d'un échantillon iid du couple (\mathbf{X}, Y) . Les estimateurs du maximum de vraisemblance (EMV) des paramètres sont :

$$\begin{aligned}\widehat{\pi}_k &= \frac{n_k}{n} \\ \widehat{\boldsymbol{\mu}}_k &= \frac{1}{n_k} \sum_{i=1}^n t_{ik} \mathbf{x}_i \\ \widehat{\Sigma}_k &= \frac{1}{n_k} \sum_{i=1}^n t_{ik} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)',\end{aligned}$$

pour $k = 1, \dots, c$, t_{ik} étant une variable binaire indiquant l'appartenance à la classe ω_k : $t_{ik} = 1$ si $y_i = \omega_k$, $t_{ik} = 0$ sinon. Comme on l'a vu dans le chapitre précédent, $\widehat{\boldsymbol{\mu}}_k$ est un estimateur sans biais de $\boldsymbol{\mu}_k$, mais l'estimateur $\widehat{\Sigma}_k$ est biaisé : on le remplace souvent par l'estimateur sans biais $S_k = \frac{n_k}{n_k - 1} \widehat{\Sigma}_k$.

La méthode consistant à remplacer, dans le modèle précédent, les paramètres par leurs EMV (éventuellement corrigés) est appelée *analyse discriminante quadratique* (ADQ).

3 Analyse discriminante linéaire

3.1 Modèle

On suppose cette fois que la matrice de variance est commune à toute les classes (hypothèse d'homoscédasticité) : $\Sigma_k = \Sigma$, $k \in \{1, \dots, c\}$. On a donc

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} (\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

En calculant $\ln(\pi_k f_k(\mathbf{x}))$ et en supprimant les termes identiques pour toutes les classes, on obtient les fonctions discriminantes suivantes :

$$g_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) + \ln \pi_k. \quad (3)$$

Le terme $(\mathbf{x} - \boldsymbol{\mu}_k)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)$ est le carré de la *distance de Mahalanobis* de \mathbf{x} à $\boldsymbol{\mu}_k$. Lorsque les probabilités a priori sont égales, la règle de Bayes avec coûts $\{0, 1\}$ revient donc à affecter l'individu à la classe dont le centre est le plus proche de \mathbf{x} , au sens de la distance de Mahalanobis.

En développant le membre de droite de (3) et en remarquant que le terme quadratique ne dépend pas de k , on obtient les nouvelles fonctions discriminantes suivantes :

$$h_k(\mathbf{x}) = (\Sigma^{-1} \boldsymbol{\mu}_k)' \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k' \Sigma^{-1} \boldsymbol{\mu}_k + \ln \pi_k.$$

Ces fonctions discriminantes sont linéaires : la règle de Bayes est donc dans ce cas une *règle de décision linéaire*.

Les régions de décision \mathcal{R}_k^* et \mathcal{R}_ℓ^* sont séparées par une frontière d'équation :

$$h_k(\mathbf{x}) = h_\ell(\mathbf{x}) \Leftrightarrow (\Sigma^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell))' \left(\mathbf{x} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell}{2} + \frac{\ln(\pi_k/\pi_\ell)}{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)' \Sigma^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \right) = 0.$$

C'est un hyperplan de vecteur normal $\Sigma^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)$. Si $\pi_k = \pi_\ell$, cet hyperplan passe par le centre du segment d'extrémités $\boldsymbol{\mu}_k$ et $\boldsymbol{\mu}_\ell$.

3.2 Estimation des paramètres

Les paramètres du modèles sont les π_k , $\boldsymbol{\mu}_k$ ($k = 1, \dots, c$) et la matrice de variance Σ commune aux c classes. Les estimateurs du maximum de vraisemblance de ces paramètres sont :

$$\begin{aligned} \hat{\pi}_k &= \frac{n_k}{n} \\ \hat{\boldsymbol{\mu}}_k &= \frac{1}{n_k} \sum_{i=1}^n t_{ik} \mathbf{x}_i \\ \hat{\Sigma} &= \frac{1}{n} \sum_{k=1}^c \sum_{i=1}^n t_{ik} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)' \\ &= \frac{1}{n} \sum_{k=1}^c (n_k - 1) S_k. \end{aligned}$$

On montre que $\mathbb{E}(\hat{\Sigma}) = \frac{n-c}{n} \Sigma$. On utilise donc plutôt l'estimateur sans biais :

$$S = \frac{1}{n-c} \sum_{k=1}^c (n_k - 1) S_k.$$

La méthode consistant à remplacer, dans le modèle précédent, les paramètres par leurs EMV (éventuellement corrigés) est appelée *analyse discriminante linéaire* (ADL).

4 Autres modèles

4.1 Hypothèse d'indépendance conditionnelle

Il est possible de définir plusieurs variantes des modèles précédents en faisant différentes hypothèses sur les matrices de variance. Par exemple, une hypothèse courante consiste à supposer l'indépendance des p variables conditionnellement à Y , ce qui, dans le modèle gaussien, revient à supposer les matrices Σ_k diagonales. On parle quelquefois de classifieur bayésien *naïf*. Si l'on fait cette hypothèse, on obtient une variante de l'ADQ dans laquelle les matrices de variance Σ_k sont estimées par $\text{diag}(S_k)$ (on annule, dans la matrice S_k , les termes non diagonaux).

On peut également conjuguer cette hypothèse avec celle d'homoscédasticité : dans ce cas, on obtient une variante de l'ADL dans laquelle la matrice de variance commune Σ est estimée par $\text{diag}(S)$ (on annule, dans la matrice S , les termes non diagonaux).

4.2 Classifieur euclidien

Il s'agit du modèle le plus simple. On suppose que :

- les matrices de variance sont scalaires et communes à toutes les classes : on a donc $\Sigma_k = \sigma^2 I_p$, où σ^2 est la variance commune des p variables conditionnellement à chaque classe, et I_p est la matrice identité d'ordre p ;
- les probabilités a priori sont égales : $\pi_k = 1/c$, $k = 1, \dots, c$.

Dans ce cas, les densités conditionnelles ont pour expression

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \sigma^p} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu}_k)'(\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

En calculant $\ln(\pi_k f_x(\mathbf{x}))$ et en supprimant les termes identiques pour toutes les classes, on obtient les fonctions discriminantes suivantes :

$$g_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'(\mathbf{x} - \boldsymbol{\mu}_k). \quad (4)$$

Le terme $(\mathbf{x} - \boldsymbol{\mu}_k)'(\mathbf{x} - \boldsymbol{\mu}_k)$ est le carré de la *distance euclidienne* de \mathbf{x} à $\boldsymbol{\mu}_k$. La règle de Bayes avec coûts $\{0, 1\}$ revient donc dans ce cas à affecter l'individu à la classe dont le centre est le plus proche de \mathbf{x} , au sens de la distance euclidienne.

En développant le membre de droite de (4) et en remarquant que le terme quadratique ne dépend pas de k , on obtient les nouvelles fonctions discriminantes linéaires suivantes :

$$h_k(\mathbf{x}) = \boldsymbol{\mu}'_k \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}'_k \boldsymbol{\mu}_k.$$

Les régions de décision \mathcal{R}_k^* et \mathcal{R}_ℓ^* sont séparées par une frontière d'équation :

$$h_k(\mathbf{x}) = h_\ell(\mathbf{x}) \Leftrightarrow (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)' \left(\mathbf{x} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell}{2} \right) = 0.$$

C'est l'hyperplan médiateur du segment d'extrémités $\boldsymbol{\mu}_k$ et $\boldsymbol{\mu}_\ell$.

Notons que cette règle ne dépend que des moyennes $\boldsymbol{\mu}_k$, qui peuvent être estimées par $\hat{\boldsymbol{\mu}}_k$. Le classifieur correspondant est appelé *classifieur euclidien*.

4.3 Comparaison des différents modèles

Les différents modèles étudiés dans ce chapitre diffèrent par leurs nombres de paramètres, comme indiqué dans le tableau 1. A priori, il pourrait sembler souhaitable de faire le moins d'hypothèses possible et de se placer d'emblée dans le cas le plus général. Cependant, il s'avère que, lorsque le nombre de paramètres à estimer est trop important, les erreurs d'estimation compensent le gain de flexibilité du modèle. Il faut donc, en pratique, réaliser un compromis et rechercher un modèle de complexité adaptée à la taille de l'ensemble d'apprentissage. Nous étudierons, dans le prochain chapitre, des méthodes de *sélection de modèle* permettant de choisir le meilleur classifieur parmi un ensemble de règles de décision.

4.4 Analyse discriminante régularisée (ADR)

Cette méthode permet de définir une infinité de règles de décision intermédiaires entre l'ADQ et l'ADL. Posons

$$\hat{\Sigma}_k(\lambda) = \frac{(1-\lambda)(n_k-1)S_k + \lambda(n-c)S}{(1-\lambda)(n_k-1) + \lambda(n-c)},$$

TAB. 1 – Nombres de paramètres associés aux différents modèles.

| Modèle | Nombre de paramètres |
|--------------------------------------|---|
| ADQ | $c \left(p + \frac{p(p+1)}{2} \right) + c - 1$ |
| ADQ avec indépendance conditionnelle | $2cp + c - 1$ |
| ADL | $cp + \frac{p(p+1)}{2} + c - 1$ |
| ADL avec indépendance conditionnelle | $cp + p + c - 1$ |
| Classifieur euclidien | cp |

avec $\lambda \in [0, 1]$. Si $\lambda = 1$, on a $\widehat{\Sigma}_k(\lambda) = S$ et on retrouve l'ADL. Si $\lambda = 0$, on a $\widehat{\Sigma}_k(\lambda) = S_k$ et on retrouve l'ADQ. Pour $0 < \lambda < 1$, on a une infinité de solutions intermédiaires.

Si n est inférieur ou comparable à p , l'approche précédente, restant intermédiaire entre l'ADL et l'ADQ en termes de complexité, peut donner de moins bons résultats qu'un modèle plus simple comme l'ADL avec hypothèse d'indépendance conditionnelle, ou le classifieur euclidien. Une variante consiste donc à poser

$$\widehat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma)\widehat{\Sigma}_k(\lambda) + \gamma c_k I_p,$$

avec

$$c_k = \frac{\text{tr } \widehat{\Sigma}_k(\lambda)}{p}.$$

Pour $\gamma = 0$, $\widehat{\Sigma}_k(\lambda, \gamma)$ est identique à l'estimateur $\widehat{\Sigma}_k(\lambda)$ précédent. Pour $\gamma = 1$, $\widehat{\Sigma}_k(\lambda, \gamma)$ est une matrice scalaire, ce qui revient à supposer que les p variables sont indépendantes conditionnellement à Y , et qu'elles ont la même variance conditionnelle.

A chaque valeur donnée à γ et λ correspond un estimateur des matrices de variance conditionnelles, et donc un nouveau classifieur lorsqu'on injecte ces estimateurs dans l'expression des fonctions discriminantes (2). Se pose donc le problème du choix de ces *hyperparamètres*. Ce problème sera abordé dans le chapitre suivant.

5 Probabilité d'erreur de Bayes

5.1 Expression exacte ($c = 2$, $\Sigma_k = \Sigma$)

Dans certains cas simples, il est possible de calculer exactement la probabilité d'erreur de Bayes. Dans ce paragraphe, nous nous placerons dans le cas de deux classes, avec hypothèse d'homoscédasticité.

Dans ce cas, la règle de Bayes avec coûts $\{0, 1\}$ peut s'écrire

$$g^*(\mathbf{x}) = \begin{cases} a_1 & \text{si } h(\mathbf{x}) < \ln \frac{\pi_1}{\pi_2} \\ a_2 & \text{sinon,} \end{cases}$$

avec

$$h(\mathbf{x}) = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right)' \Sigma^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

On montre que

$$h(\mathbf{X}) \stackrel{\omega_1}{\approx} \mathcal{N}\left(-\frac{\Delta^2}{2}, \Delta^2\right),$$

et

$$h(\mathbf{X}) \stackrel{\omega_2}{\approx} \mathcal{N}\left(\frac{\Delta^2}{2}, \Delta^2\right),$$

avec $\Delta^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \Sigma^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. La quantité Δ^2 est appelée *carré de la distance de Mahalanobis* entre les deux classes.

On en déduit l'expression de la probabilité d'erreur de la règle de Bayes :

$$\epsilon^* = \mathbb{P}\left(h(\mathbf{X}) < \ln \frac{\pi_1}{\pi_2} | \omega_2\right) \pi_2 + \mathbb{P}\left(h(\mathbf{X}) \geq \ln \frac{\pi_1}{\pi_2} | \omega_1\right) \pi_1 \quad (5)$$

$$= \phi\left(\frac{\ln(\pi_1/\pi_2) - \Delta^2/2}{\Delta}\right) \pi_2 + \left[1 - \phi\left(\frac{\ln(\pi_1/\pi_2) + \Delta^2/2}{\Delta}\right)\right] \pi_1. \quad (6)$$

Dans le cas $\pi_1 = \pi_2$, on a donc

$$\epsilon^* = \phi\left(-\frac{\Delta}{2}\right).$$

5.2 Borne de Bhattacharyya

Dans le cas général, même en se limitant à $c = 2$, il n'est pas possible d'exprimer analytiquement l'erreur de Bayes. Cependant, on peut en donner des approximations.

Dans le cas de deux classes, on a

$$\epsilon^* = \int_{\mathbb{R}^p} \min(\mathbb{P}(\omega_1|\mathbf{x}), \mathbb{P}(\omega_2|\mathbf{x})) f(\mathbf{x}) d\mathbf{x} \quad (7)$$

$$= \int_{\mathbb{R}^p} \min(f_1(\mathbf{x})\pi_1, f_2(\mathbf{x})\pi_2) d\mathbf{x}. \quad (8)$$

Or, $\min(a, b) \leq \sqrt{ab}$ pour tous réels positifs a et b . On en déduit une borne supérieure de l'erreur de Bayes :

$$\epsilon^* \leq \sqrt{\pi_1 \pi_2} \int_{\mathbb{R}^p} \sqrt{f_1(\mathbf{x}) f_2(\mathbf{x})} d\mathbf{x} = \sqrt{\pi_1 \pi_2} e^{-\Delta_B^2/8}.$$

La quantité Δ_B^2 est appelée *carré de la distance de Bhattacharyya* entre les deux classes. Dans le cas gaussien, on montre qu'elle est égale à :

$$\Delta_B^2 = \frac{1}{8} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2} \ln \frac{\det \frac{\Sigma_1 + \Sigma_2}{2}}{\sqrt{\det \Sigma_1 \det \Sigma_2}}.$$

Cette quantité est donc composée de deux termes, dont le premier dépend de la différence des moyennes, et le second de la différence des variances. La distance de Bhattacharyya est souvent utilisée comme mesure de distance entre deux classes, même en dehors de l'hypothèse gaussienne (mais son interprétation liée à une borne supérieure de l'erreur de Bayes n'est alors plus valide).