Apprentissage statistique : Analyse discriminante

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)

HMMA303

1. Introduction
2. Cas paramétrique : modélisation gaussienne des distributions dans les classes
3. Analyse discriminante linéaire
   - Estimation de π_0 et π_1
   - Estimation de μ_0 et μ_1
   - Estimation de Σ
   - Prédiction à partir de $\mathbb{P}_{\hat{\theta}}$
4. Analyse discriminante quadratique
5. Analyse discriminante non paramétrique
   - Méthode du noyau
     - Approximation en dimension 1
     - Approximation en dimension supérieure
   - Application à l'estimation de $\mathbb{P}(Y = 1|X = x)$

Introduction

On vise ici l'estimation de $\mathbb{P}(Y = 0|X = x)$ et $\mathbb{P}(Y = 1|X = x)$

Une fois cette estimation effectuée, on peut aisément construire un classifieur et l'on dispose d'un score

L'analyse discriminante linéaire ou quadratique s'applique au cas où $x \in \mathbb{R}^d$, les d variables prédictives sont quantitatives

Ces deux méthodes sont paramétriques, elles postulent que le couple (X, Y) appartient à une famille de lois $(\mathbb{P}_\theta)_{\theta \in \Theta}$
où $\Theta \subseteq \mathbb{R}^p$ est un espace de dimension finie

L'analyse discriminante non paramétrique ne fait pas cette hypothèse, l'objet à estimer appartient à un espace de dimension infinie (typiquement une fonction)

Cas paramétrique : modélisation gaussienne des distributions dans les classes

$y \in \{0, 1\}$ et $x \in \mathbb{R}^d$

Hypothèses
$X|Y = 0 \sim \mathcal{N}(\mu_0, \Sigma_0)$
$X|Y = 1 \sim \mathcal{N}(\mu_1, \Sigma_1)$

$\mu_0 \in \mathbb{R}^d$, $\mu_1 \in \mathbb{R}^d$, Σ_0 et Σ_1 sont des matrices de covariance que l'on suppose non dégénérées (déterminants strictement positifs)

Cas paramétrique : modélisation gaussienne des distributions dans les classes

Dans ce cas,

$$f_j(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right\}$$

Notons $\pi_0 = \mathbb{P}(Y=0)$ et $\pi_1 = \mathbb{P}(Y=1)$

Cas paramétrique : modélisation gaussienne des distributions dans les classes

On dispose d'un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ du couple (X, Y)

On allons estimer $\theta = (\mu_0, \Sigma_0, \mu_1, \Sigma_1, \pi_0)$ dans le but d'estimer $\mathbb{P}(Y = 1|X = x)$

D'après la règle de Bayes

$$\hat{\mathbb{P}}(Y = 1|X = x) = \frac{\hat{\pi}_1 f_1(x; \hat{\mu}_1, \hat{\Sigma}_1)}{\hat{\pi}_0 f_0(x; \hat{\mu}_0, \hat{\Sigma}_0) + \hat{\pi}_1 f_1(x; \hat{\mu}_1, \hat{\Sigma}_1)}$$

Analyse discriminante linéaire

Dans le cas LDA (Linear Discriminant Analysis), on restreint le modèle à des gaussiennes de même matrice de covariance $\Sigma_0 = \Sigma_1 = \Sigma$

On doit alors estimer $1 + 2d + \frac{d(d-1)}{2} + d = \frac{5d}{2} + \frac{d^2}{2} + 1$

Plusieurs méthodes d'estimation

- maximum de vraisemblance
- méthode des moments
- méthodes d'estimation robustes

Analyse discriminante linéaire
Estimation de π_0 et π_1

On note $N_0 = \sum_{i=1}^{n} \mathbf{1}_{Y_i=0}$ et $N_1 = \sum_{i=1}^{n} \mathbf{1}_{Y_i=1} = n - N_0$

$\hat{\pi}_0 = \frac{N_0}{n}$ et $\hat{\pi}_1 = \frac{N_1}{n}$, $N_0 + N_1 = n$

Ces estimateurs sont extrêmement intuitifs, ils sont sans biais et de variance minimale

Cependant, ils sont basés sur l'hypothèse selon laquelle notre échantillon est représentatif de la population totale

Analyse discriminante linéaire
Estimation de π_0 et π_1

Dans de nombreux cas ce n'est pas vrai pour la loi marginale de Y, notamment si l'on choisit les individus sur lesquels l'analyse est réalisée

Dans un tel cas, on n'estime pas π_j par $\hat{\pi}_j$ mais par une estimation issue d'une étude préalable ou une estimation fournie par des experts

Analyse discriminante linéaire
Estimation de μ_0 et μ_1

Le maximum de vraisemblance et la méthode des moments donnent les mêmes estimateurs

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{i=1}^{n} X_i \mathbf{1}_{Y_i=j}$$

Ces estimateurs sont extrêmement intuitifs, ils sont sans biais et de variance minimale

On estime l'espérance mathématique par la moyenne empirique

Analyse discriminante linéaire
Estimation de Σ

On utilise classiquement l'estimateur par la méthode des moments qui est sans biais contrairement à l'estimateur du maximum de vraisemblance

$$\hat{\Sigma} = \frac{1}{n-2} \left\{ \sum_{i=1}^{n} \mathbf{1}_{Y_i=0} (X_i - \hat{\mu}_0)(X_i - \hat{\mu}_0)^T + \sum_{i=1}^{n} \mathbf{1}_{Y_i=1} (X_i - \hat{\mu}_1)(X_i - \hat{\mu}_1)^T \right\}$$

Analyse discriminante linéaire
Prédiction à partir de $\mathbb{P}_{\hat{\theta}}$

Sur la base de l'estimation de θ par $\hat{\theta}$, on estime $\mathbb{P}(Y = 1|X = x)$ par

$$\frac{N_1 \exp\left\{-\frac{1}{2}(x - \hat{\mu}_1)^T \hat{\Sigma}^{-1}(x - \hat{\mu}_1)\right\}}{N_0 \exp\left\{-\frac{1}{2}(x - \hat{\mu}_0)^T \hat{\Sigma}^{-1}(x - \hat{\mu}_0)\right\} + N_1 \exp\left\{-\frac{1}{2}(x - \hat{\mu}_1)^T \hat{\Sigma}^{-1}(x - \hat{\mu}_1)\right\}}$$

Si l'on singe le classifieur de Bayes, on affecte le point x à la classe 1 si $\hat{\mathbb{P}}(Y = 1|X = x) \geqslant 1/2$

Analyse discriminante linéaire
Prédiction à partir de $\mathbb{P}_{\hat{\theta}}$

L'ensemble des points affectés à la classe 1 est

$$\{x \in \mathbb{R}^d | (\hat{\mu}_0 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} x \geqslant b\}$$

où b est un scalaire dépendant de $\hat{\theta}$

On remarque que la frontière $(\hat{\mu}_0 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} x = b$ est un hyperplan : c'est pourquoi on parle d'analyse discriminante linéaire

Analyse discriminante quadratique

On ne contraint plus les structures de covariance à être identiques $\Sigma_0 \neq \Sigma_1$

On doit alors estimer $1 + 2d + d(d-1) + 2d = d^2 + 3d + 1$

Les paramètres π_0, π_1, μ_0 et μ_1 sont estimés comme dans le cas linéaire

Analyse discriminante quadratique

Par la méthode des moments, l'estimateur de Σ_0 est

$$\hat{\Sigma}_0 = \frac{1}{N_0 - 1} \sum_{i=1}^{n} \mathbf{1}_{Y_i=0} (X_i - \hat{\mu}_0)(X_i - \hat{\mu}_0)^T$$

De même, Σ_1 est estimé par

$$\hat{\Sigma}_1 = \frac{1}{N_1 - 1} \sum_{i=1}^{n} \mathbf{1}_{Y_i=1} (X_i - \hat{\mu}_1)(X_i - \hat{\mu}_1)^T$$

On en déduit une estimation de $\mathbb{P}(Y = 1|X = x)$

Analyse discriminante quadratique

Si l'on singe le classifieur de Bayes, on affecte le point x à la classe 1 si $\hat{\mathbb{P}}(Y = 1|X = x) \geqslant 1/2$

L'ensemble des points affectés à la classe 1 est tel que

$$\left\{ x \in \mathbb{R}^d | -\frac{1}{2}(x - \hat{\mu}_1)^T \hat{\Sigma}_1^{-1}(x - \hat{\mu}_1) + \frac{1}{2}(x - \hat{\mu}_0)^T \hat{\Sigma}_0^{-1}(x - \hat{\mu}_0) \geqslant d' \right\}$$

La frontière séparant les 2 classes est une hypersurface quadratique

Analyse discriminante non paramétrique

En statistique, on parle d'estimation non paramétrique ou fonctionnelle lorsque le nombre de paramètres à estimer est infini

Par exemple, au lieu de supposer que l'on a affaire à une densité de type connu (dont on estime les paramètres), on cherche à estimer la fonction de densité toute entière

Pour tout $x \in E$, $f(x)$ est estimée par $\hat{f}(x)$

Approche très souple mais qui n'est applicable qu'avec des échantillons de grande taille

Analyse discriminante non paramétrique

$Y \in \{0, 1\}$, $\pi_0 = \mathbb{P}(Y = 0)$, $\pi_1 = \mathbb{P}(Y = 1)$ et
$f_j(x)$ densité de $X|Y = j$

On ne suppose plus que les $f_j(x)$ sont gaussiennes mais on les estime totalement

Analyse discriminante non paramétrique
Méthode du noyau

On considère avant tout le cas de variables prédictives quantitatives $X \in \mathbb{R}^d$

Definition

On appelle noyau de probabilité sur $\mathbb{R}^d$, une fonction $K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ telle que $K(\cdot, x_0)$ est une densité de probabilité

Definition

L'approximation noyau de f (relativement au noyau K) est

$$\hat{f}_K(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, X_i)$$

où $X_1, \ldots, X_n$ est un n-échantillon de X

Analyse discriminante non paramétrique
Méthode du noyau

Pour construire un noyau, on part d'une densité de probabilité sur $\mathbb{R}$, $m(x)$, symétrique

- $m(x) \geqslant 0$ pour tout $x \in \mathbb{R}$
- $\int_{\mathbb{R}} m(x)dx = 1$
- $m(-x) = m(x)$ pour tout $x \in \mathbb{R}$

Puis, on définit

$$K(x, x') = \frac{1}{h} m\left(\frac{x - x'}{h}\right)$$

avec $h > 0$

Analyse discriminante non paramétrique
Méthode du noyau

Example (noyau rectangulaire)

$$m(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$$

$$K_h(x, x') = \frac{1}{h} m\left(\frac{x - x'}{h}\right) = \frac{1}{2h}\mathbf{1}_{[x'-h,x'+h]}(x)$$

Pour $\sigma = 1$, on estime $f(x)$ par

$$\hat{f}(x) = \frac{1}{2n}\sum_{i=1}^{n} \mathbf{1}_{[x-1,x+1]}(X_i)$$

Analyse discriminante non paramétrique
Méthode du noyau

Example (noyau triangulaire)

$$m(x) = (1 - |x|)\mathbf{1}_{[-1,1]}(x)$$

$$K_h(x, x') = \frac{1}{h}\left(1 - \left|\frac{x - x'}{h}\right|\right)\mathbf{1}_{[x'-h,x'+h]}(x)$$

Analyse discriminante non paramétrique
Méthode du noyau

Example (noyau gaussien)

$$m(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

$$K_h(x, x') = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{(x - x')^2}{2h^2}\right)$$

h c'est la fenêtre, plus h est petit plus l'estimation est versatile et le modèle sous-jacent complexe

Analyse discriminante non paramétrique
Méthode du noyau

Ayant fait le choix d'un noyau uni-dimensionnel associé à la densité $m(\cdot)$, on peut procéder par tensorisation

- mono-fenêtre $\sigma > 0$ $x, x' \in \mathbb{R}^d$

$$K_h(x, x') = \frac{1}{h^d} \prod_{j=1}^{d} m\left(\frac{x_j - x'_j}{h}\right)$$

- multi-fenêtres $h = (h_1, \ldots, h_d)$, $h_i > 0$

$$K_h(x, x') = \prod_{j=1}^{d} \frac{1}{h_j} m\left(\frac{x_j - x'_j}{h_j}\right)$$

Analyse discriminante non paramétrique
Méthode du noyau

Lorsque l'on estime la densité de X avec ces deux méthodes, les composantes de X sont supposées indépendantes

Difficile de calibrer un estimateur à noyaux avec des covariances

En pratique, on utilise la solution multi-fenêtres et on se ramène donc au choix d'un noyau en dimension 1 et à la calibration de d fenêtres

Analyse discriminante non paramétrique
Méthode du noyau

Typiquement, on utilise un noyau gaussien et on fixe h_j à

$$0.9 \min\left(\hat{\sigma}_j, \frac{\hat{q}_{j,0.75} - \hat{q}_{j,0,25}}{1.34}\right) n^{-1/5}$$

C'est la règle du pouce de Silverman

Analyse discriminante non paramétrique
Application à l'estimation de $\mathbb{P}(Y = 1|X = x)$

On se place dans le contexte de la classification supervisée binaire

On considère que l'on dispose de d variables prédictives dont q quantitatives et $d - q$ qualitatives

On dispose d'un n-échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$

Objectif : estimer $\mathbb{P}(Y = 1|X = x)$

Analyse discriminante non paramétrique
Application à l'estimation de $\mathbb{P}(Y = 1|X = x)$

Comme pour l'analyse discriminante linéaire/quadratique, on estime avant tout $f_0(x)$ et $f_1(x)$ densités conditionnelles de $X|Y = 0$ et $X|Y = 1$

Analyse discriminante non paramétrique
Application à l'estimation de $\mathbb{P}(Y = 1|X = x)$

Méthode naïve Bayes pour estimation de $f_j(x)$

- on prend les x_i tels que $y_i = j$, $x_i \in \mathbb{R}^d$
- on estime par la méthode à noyau multi-fenêtres la densité des q variables quantitatives
- on estime la densité des $d - q$ variables qualitatives en utilisant les fréquences empiriques correspondantes
- on en déduit $\hat{f}_j(x)$

Analyse discriminante non paramétrique
Application à l'estimation de $\mathbb{P}(Y = 1|X = x)$

Hypothèse sous-jacente importante : les variables prédictives sont supposées indépendantes conditionnellement à Y

Pour estimer π_0 et π_1 on fait comme dans le cas de l'analyse discriminante linéaire/quadratique

Formule de Bayes

$$\hat{\mathbb{P}}(Y = 1|X = x) \frac{\hat{\pi}_1 \hat{f}_1(x)}{\hat{\pi}_0 \hat{f}_0(x) + \hat{\pi}_1 \hat{f}_1(x)}$$