Distributions normales

Thierry Chateau

Pascal Institute

2017

Plan

Introduction

- Les classifieurs Bayesiens sont basés sur des fonctions de densité de probabilité ($p(\mathbf{x}|\omega_i)$)
- Si $p(\mathbf{x}|\omega_i)$ est trop complexe, les temps de calcul sont élevés (intégrales présentes dans la formule de Bayes)
- D'où l'idée de modéliser $p(\mathbf{x}|\omega_i)$ par une fonction analytique définie avec seulement quelques paramètres
- La fonction la plus souvent utilisée est une loi normale (ou loi de Gauss) :
  - utilisée dans de nombreuses applications,
  - permet de modéliser de manière naturelle des caractéristiques bruitées aléatoirement
  - s'applique lorsque, pour une classe, les vecteurs $\mathbf{x}$ sont répartis de manière continue et "harmonieuse" autour d'un vecteur moyen μ

Rappels : espérance, variance

- Soit $p(x)$ une densité de probabilités et $f(x)$ une fonction scalaire
- L'espérance de $f(x)$, pour la densité de probabilités $p(x)$ est définie par :

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)dx$$

Rappels : espérance, variance

- L'espérance de x est appelée **moyenne** :

$$\mu = E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

- L'espérance de la déviation quadratique $(x - \mu)^2$ est appelée variance :

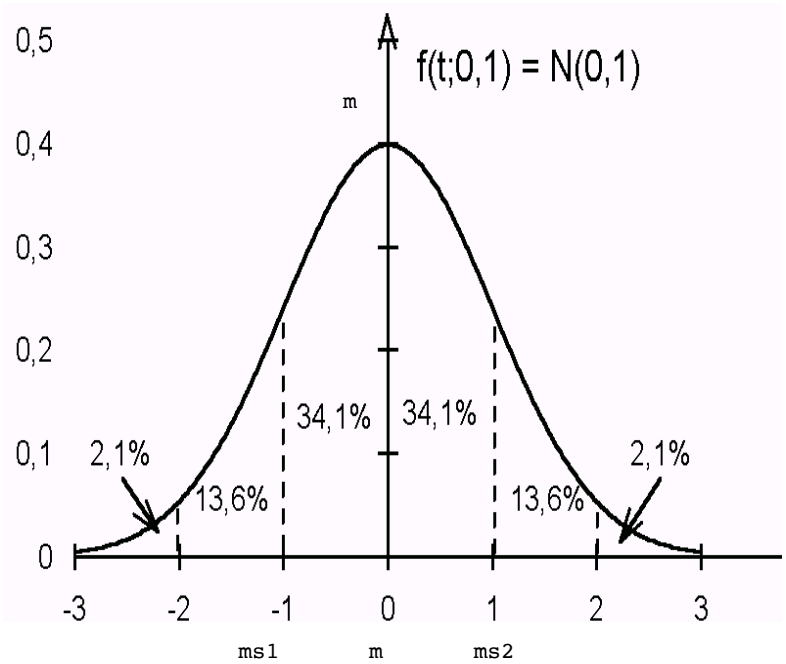
$$= E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx$$

Loi normale de dimension 1

Loi normale de dimension 1, notée $N(\mu, \sigma^2)$:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- μ : moyenne
- σ^2 : variance
- Max. de la courbe : $p(\mu) = 1/\sqrt{2\pi}\sigma$



Loi normale multivariée

- Loi normale généralisée à d dimensions et notée $N(\boldsymbol{\mu}, \Sigma)$:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$\boldsymbol{\mu}$: vecteur moyen

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\mu_i = E[x_i]$$

Loi normale multivariée

Σ : matrice de covariance

$$\Sigma = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']$$

$$\Sigma = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' p(\mathbf{x}) dx$$

$$\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$$

Dans le cas discret :

$$\sigma_{ij} = \frac{1}{N} \sum_x (x_i - \mu_i)(x_j - \mu_j)$$

La matrice de covariance

- Σ contient toute l'information sur la façon dont se répartissent les formes de la classe dans l'espace des paramètres.
- Σ décrit la dispersion des données
- Σ est symétrique $(\sigma_{ij} = \sigma_{ji})$
- si les composantes du vecteur de paramètres sont statistiquement indépendantes (x_i et x_j), alors $\sigma_{ij} = \sigma_{ji} = 0$ $i \neq j$.

Distance de Mahalanobis

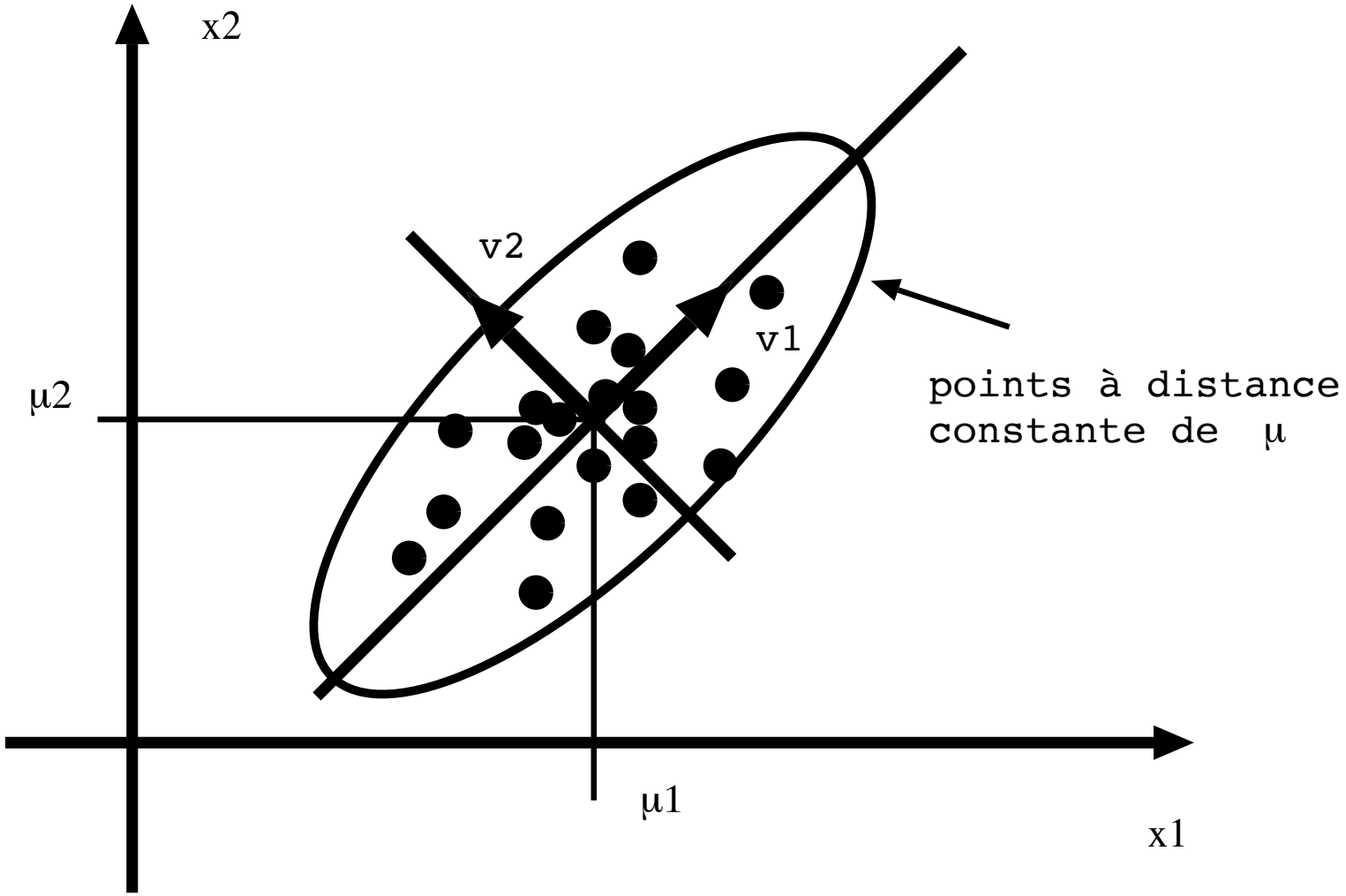
- Les courbes de densité de probabilité constantes sont modélisées par des hyperboloïdes d'équation

$$(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c, \text{ avec } c \in R$$

- Les directions des axes principaux sont définies par les valeurs propres
- La dispersion selon chaque axe principal est déterminée par la valeur propre correspondante
- La distance de Mahalanobis entre un vecteur de paramètres $\mathbf{x}$ et $\boldsymbol{\mu}$ est définie par :

$$\sqrt{(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})}$$

Exemple dans R^2



Distance de Mahalanobis

Volume de l'hyper-ellipsoide défini une distance de Mahalanobis de r :

$$V = V_d |\Sigma|^{-1/2} r^d$$

avec V_d : volume de l'hypershpère unité :

$$\text{d pair} \rightarrow V_d = \frac{\pi^{d/2}}{(d/2)}$$

$$\text{d impair} \rightarrow V_d = \frac{2^d \pi^{\frac{(d-1)}{2}} \left(\frac{d-1}{2}\right)}{d!}$$

Fonctions discriminantes

Il est possible de définir des fonctions discriminantes spécifiquement associées à une loi normale :

$$g_{\omega_i} = \log(p(\mathbf{x}|\omega_i)) + \log(P(\omega_i))$$

Dans l'hypothèse d'une loi normale pour ω_i : $N(\boldsymbol{\mu}_{\omega_i}, \Sigma_{\omega_i})$

$$g_{\omega_i} = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})^t \Sigma_{\omega_i}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i}) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\omega_i}| + \log(p(\omega_i))$$

Fonction complexe mais souvent simplifiable

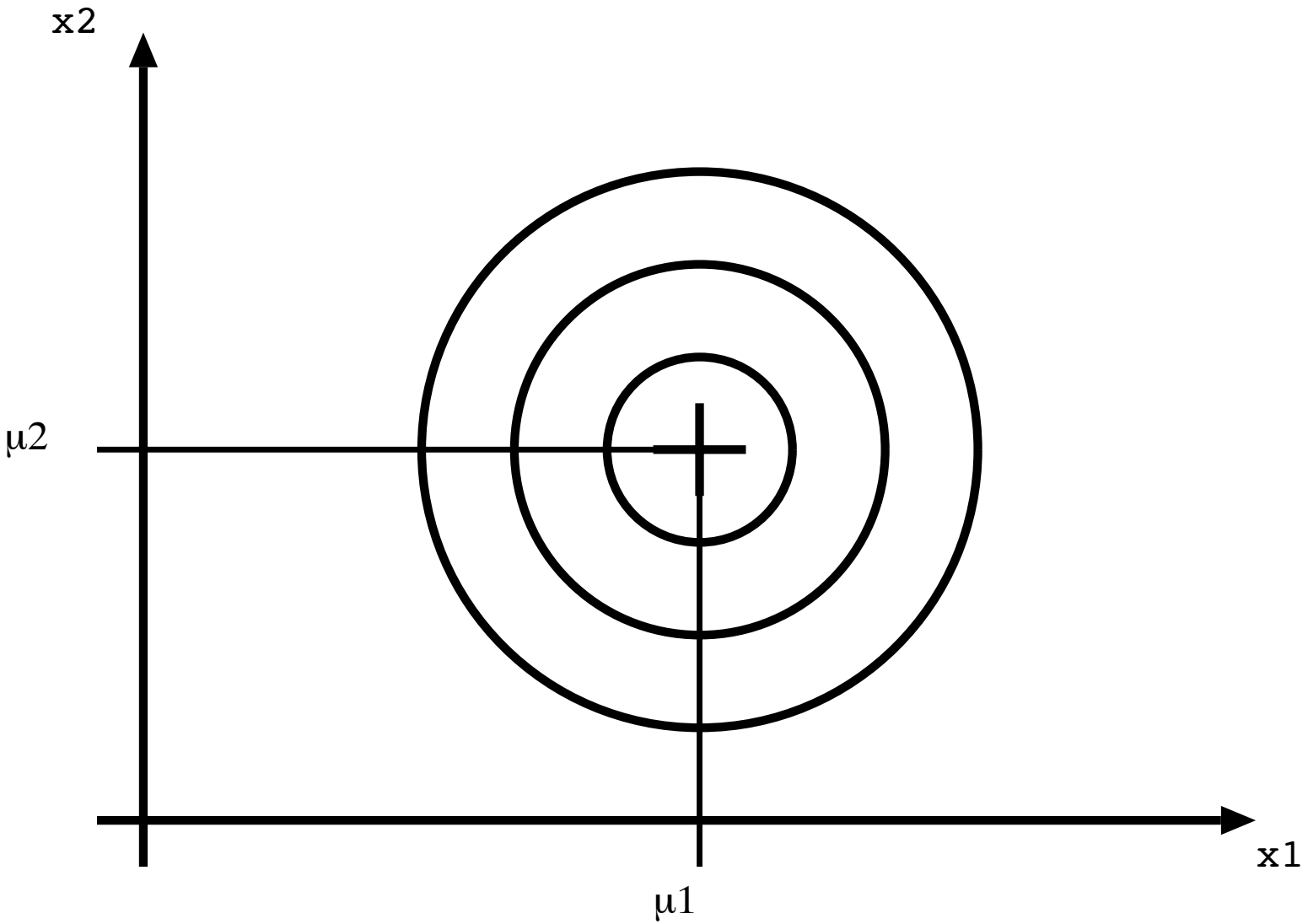
Cas 1 : $\Sigma_{\omega_i} = \sigma^2 \mathrm{I}$

$$\Sigma_{\omega_i} = \begin{pmatrix} \sigma^2 & 0 & . & 0 \\ 0 & \sigma^2 & . & 0 \\ . & . & . & 0 \\ 0 & . & 0 & \sigma^2 \end{pmatrix} \tag{1}$$

Pour toutes les classes, la variance de chaque composante du vecteur de paramètres est identique. De plus, $\mathbf{x}_i$ et $\mathbf{x}_j$ sont statistiquement indépendants.

Cas 1 : $\Sigma_{\omega_i} = \sigma^2 \mathrm{I}$

Les classes résultantes ont la forme d'une hyper-shère



Cas 1 : fonction discriminante

Les termes de la fonction dicriminante se simplifient :

$$|\Sigma_{\omega_i}| = \sigma^{2d}$$

Ce terme ne dépend plus de ω_i.
D'où la fonction discriminante :

$$g_{\omega_i} = -\frac{||\mathbf{x} - \boldsymbol{\mu}_{\omega_i}||^2}{2\sigma^2} + \log(P(\omega_i))$$

avec :

$$||\mathbf{x} - \boldsymbol{\mu}_{\omega_i}||^2 = (\mathbf{x} - \boldsymbol{\mu}_{\omega_i})^t(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})$$

carré de la distance Euclidienne

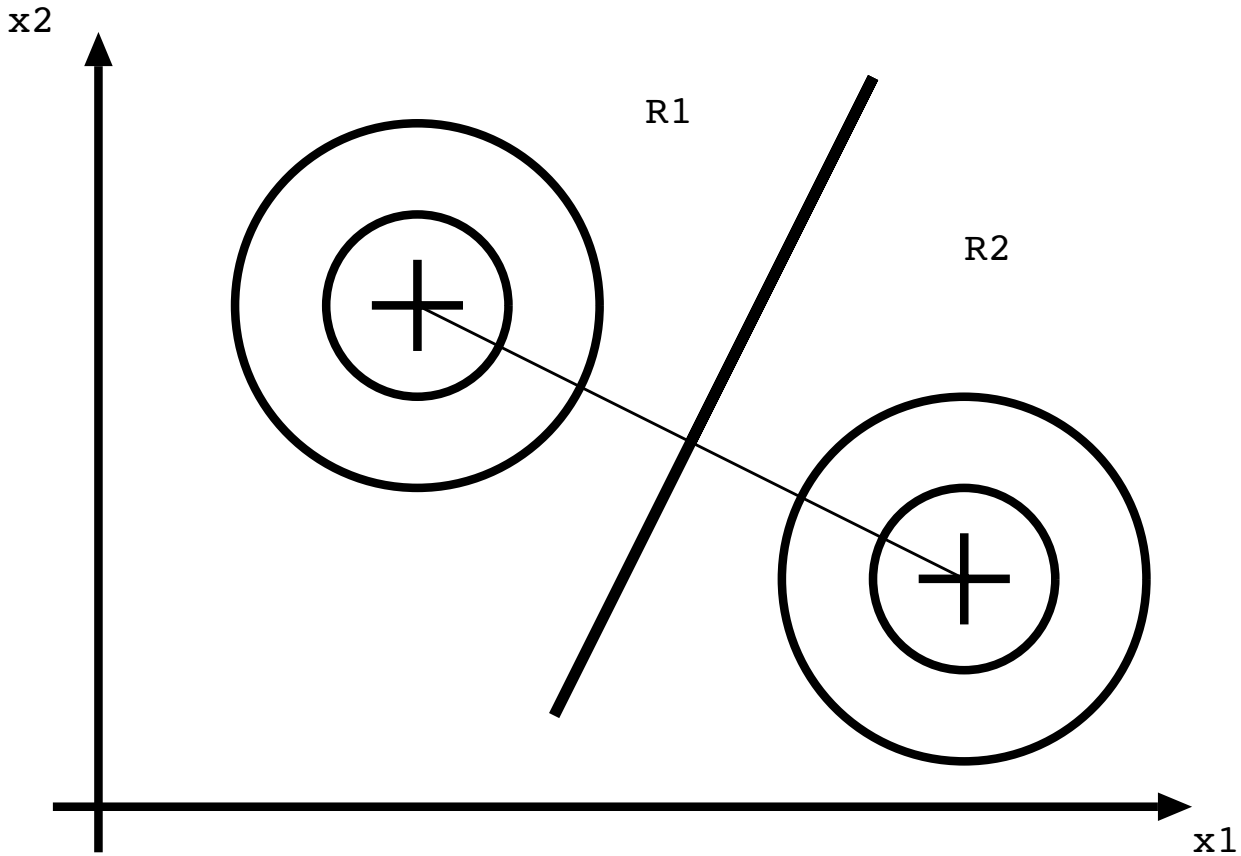
Cas 1 : fonction discriminante

Si les classes sont equiprobables :

$$g_{\omega_i} = -(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})^t(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})$$

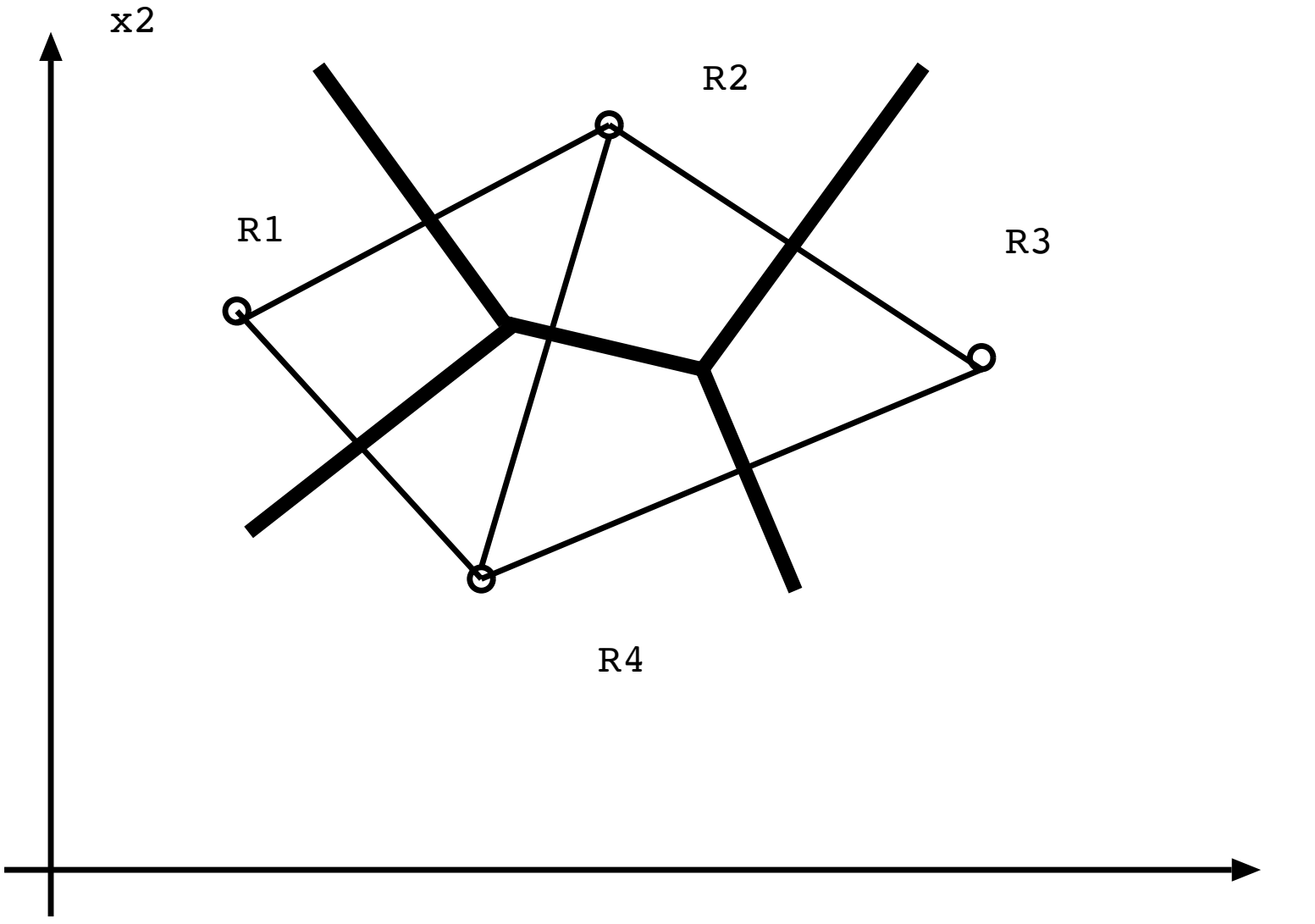
Cas 1 : règle de décision

Pour une observation **x**, la classe choisie sera celle dont le centre sera le plus proche (distance euclidienne) : **classifieur de distance minimum**



Cas 1 : règle de décision

Cas de 4 classes



Cas 1 : $P(\omega_i) \neq P(\omega_j)$

L'expression de la fonction devient :

$$g_{\omega_i}(\mathbf{x}) = \frac{1}{2\sigma}\left[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i\right] + \log(P(\omega_i))$$

On peut montrer que cela s'écrit également :

$$g_{\omega_i}(\mathbf{x}) = \mathbf{K}_i^t\mathbf{x} + K_{i0}$$

avec

$$\mathbf{K}_i = \frac{1}{\sigma^2}$$

et

$$K_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \log(P(\omega_i))$$

Cas 1 : $P(\omega_i) \neq P(\omega_j)$

$$g_{\omega_i}(\mathbf{x}) = \mathbf{K}_i^t \mathbf{x} + K_{i0}$$

C'est une fonction linéaire $\rightarrow$ les surfaces de décision sont donc des hyperplans d'équation (entre R_i et $_j$):

$$g_{\omega_i}(\mathbf{x}) = g_{\omega_j}(\mathbf{x}) \;\rightarrow\; g_{\omega_i}(\mathbf{x}) - g_{\omega_j}(\mathbf{x}) = 0$$

soit

$$(\mathbf{K}_i^t \mathbf{x} + K_{i0}) - (\mathbf{K}_j^t \mathbf{x} + K_{j0}) = 0$$

Cas 1 : $P(\omega_i) \neq P(\omega_j)$

Cette equation peut s'écrire :

$$\mathbf{W}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

avec :

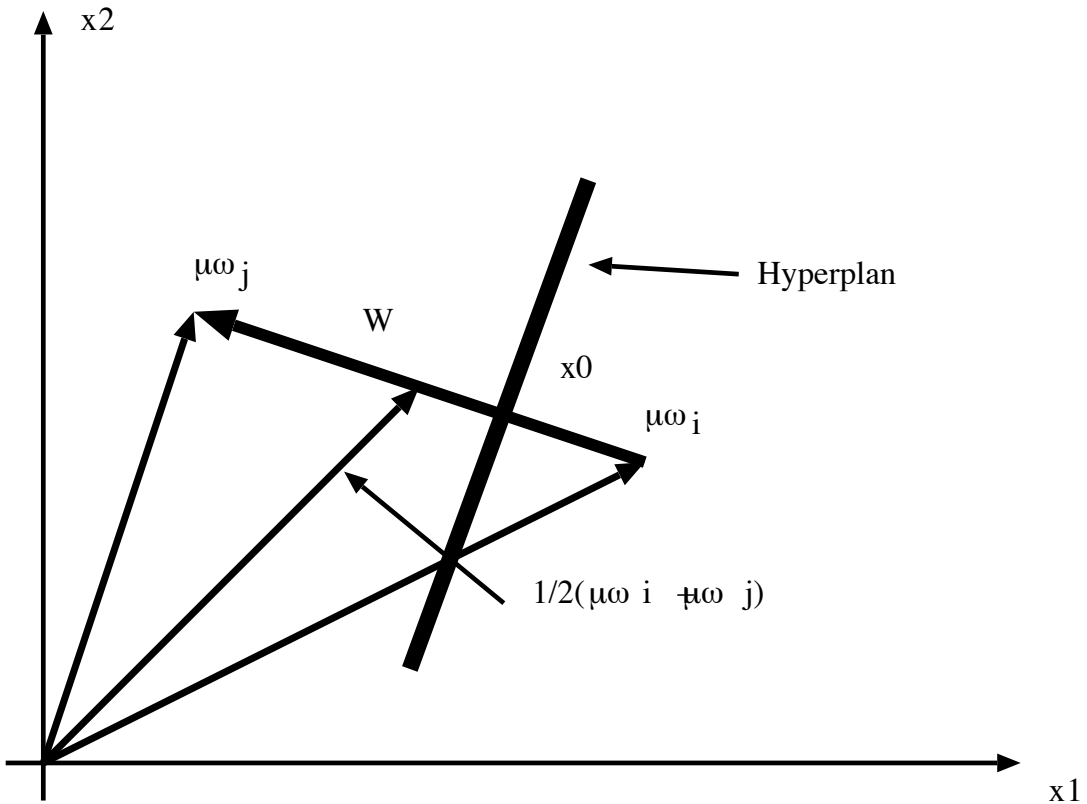
$$\mathbf{W} = \boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j}$$

et

$$\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_{\omega_i} + \boldsymbol{\mu}_{\omega_j}) - \frac{\sigma^2}{||\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j}||} \log\left(\frac{P(\omega_i)}{P(\omega_j)}\right)(\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j})$$

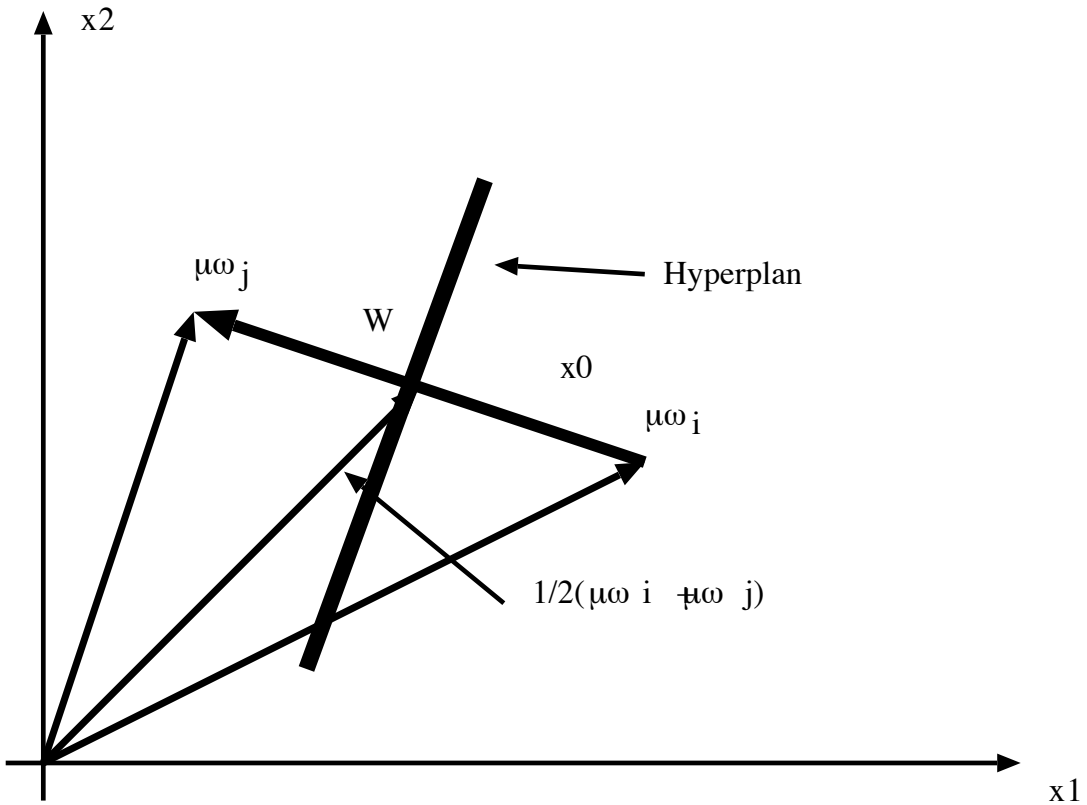
Cas 1 : $P(\omega_i) \neq P(\omega_j)$

C'est l'equation d'un hyperplan perpendiculaire à **W** et passant par $\mathbf{x}_0$



Cas 1 : $P(\omega_i) = P(\omega_j)$

$P(\omega_i) = P(\omega_j)$



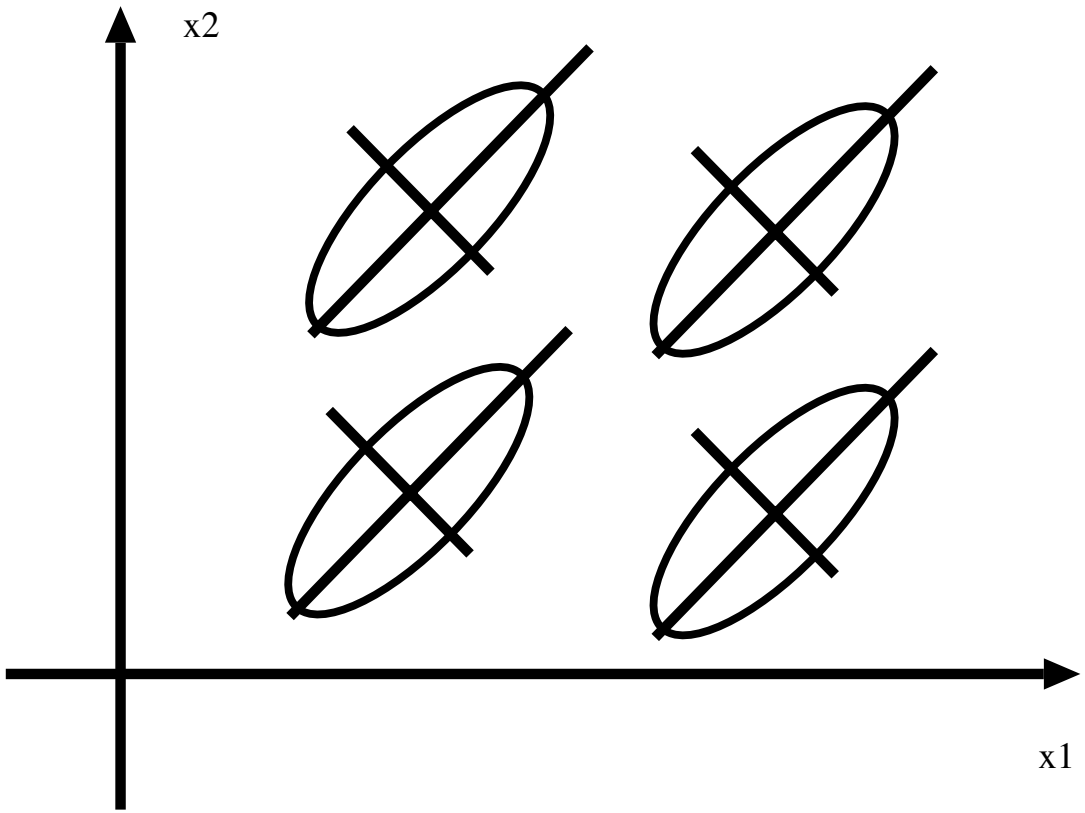
Cas 1 : $P(\omega_i) \neq P(\omega_j)$

si $\sigma^2 << ||\mathbf{mu}_{\omega_i} - \mathbf{mu}_{\omega_j}||^2$ alors $P(\omega_i)$ et (ω_j) influent peu et :

$$\mathbf{x}_0 \approx \frac{1}{2}(\mathbf{mu}_{\omega_i} - \mathbf{mu}_{\omega_j})$$

Cas 2 : $\Sigma_i = \Sigma_j$

Toutes les classes ont la même matrice de covariance :



Cas 2 : $\Sigma_i = \Sigma_j$

Fonction discriminante de départ :

$$g_{\omega_i} = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})^t \Sigma_{\omega_i}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i}) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\omega_i}| + \log(p(\omega_i))$$

Supprimons les termes indépendants de i :

$$g_{\omega_i} = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})^t \Sigma_{\omega_i}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i}) + \log(p(\omega_i))$$

avec

$$(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})^t \Sigma_{\omega_i}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})$$

distance de Mahalanobis de la forme $\mathbf{x}$ à la moyenne de la classe μ_{ω_i}

Cas 2 : $\Sigma_i = \Sigma_j$

La frontière entre deux formes est définie par : $g(\mathbf{x}) = (g_{\omega_i}(\mathbf{x}) - g_{\omega_j}(\mathbf{x})) = 0$; Soit l'equation d'un hyperplan :

$$\mathbf{W}^t(\mathbf{x} - \mathbf{x}_0) = 0$$

On obtient, après quelques calculs :

$$W = \Sigma^{-1}(\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j})$$

$$\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu}_{\omega_i} + \boldsymbol{\mu}_{\omega_j}) - \frac{\log\left(\frac{P(\omega_i)}{P(\omega_j)}\right)}{(\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j})^t \Sigma_{\omega_i}^{-1}(\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j})}(\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j})$$

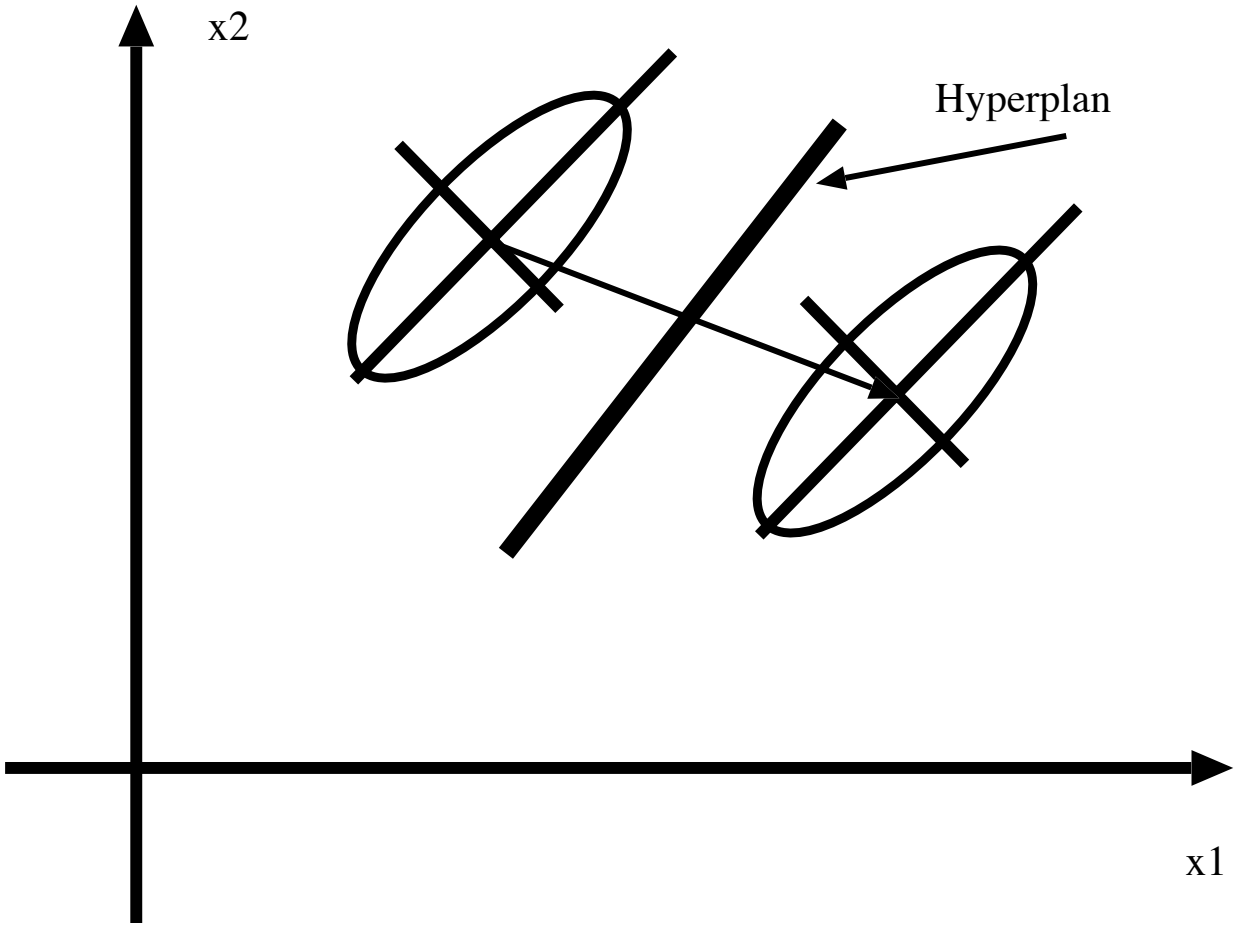
Cas 2 : $\Sigma_i = \Sigma_j$

Remarques :

- **W** n'est pas forcément dans la direction $\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j}$
- L'hyperplan séparateur n'est donc pas forcément perpendiculaire à $\boldsymbol{\mu}_{\omega_i} - \boldsymbol{\mu}_{\omega_j}$
- si $P(\omega_i) = P(\omega_j)$, l'hyperplan passe par le milieu du segment qui joint les extrémités des vecteurs $\boldsymbol{\mu}_{\omega_i}$ et $\boldsymbol{\mu}_{\omega_j}$

Cas 2 : $\Sigma_i = \Sigma_j$

$P(\omega_i) = P(\omega_j)$



Cas 3 : matrices de cov. disctinctes

Forme générale des fonctions discriminantes :

$$g_{\omega_i} = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i})^t \Sigma_{\omega_i}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\omega_i}) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\omega_i}| + \log(p(\omega_i))$$

Seul $\frac{d}{2}\log(2\pi)$ peut être éliminé. On obtient une fonction du second degrè en $\mathbf{x}$ de la forme :

$$g_{\omega_i}(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_1 \mathbf{x} + \mathbf{W}_2^t \mathbf{x} + \mathbf{W}_3$$

Cas 3 : matrices de cov. disctinctes

$$g_{\omega_i}(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_1 \mathbf{x} + \mathbf{W}_2^t \mathbf{x} + \mathbf{W}_3$$

avec :

$$\mathbf{W_1} = -\frac{1}{2}\Sigma_{\omega_i}^{-}1$$

$$\mathbf{W_2} = \Sigma_{\omega_i}^{-}1\boldsymbol{\mu}_{\omega_i}$$

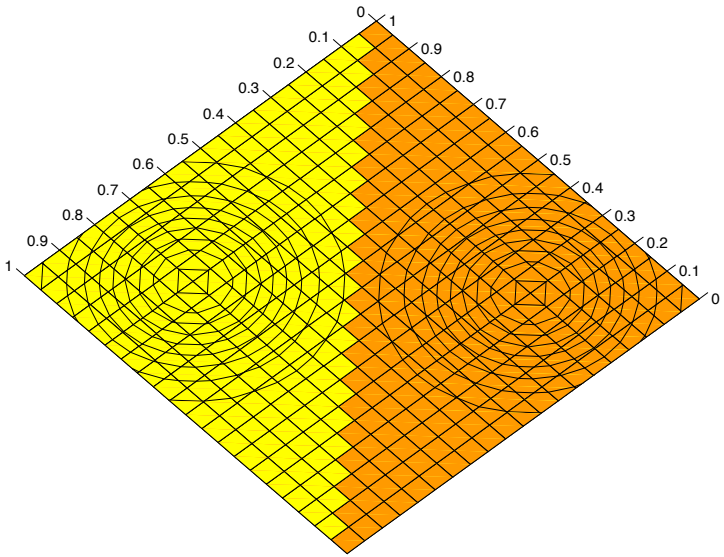
$$\mathbf{W_3} = -\frac{1}{2}\boldsymbol{\mu}_{\omega_i}^t \Sigma_{\omega_i}^{-}1\boldsymbol{\mu}_{\omega_i} - \frac{1}{2}\log\left(|\Sigma_{\omega_i}|\right) + \log(P(\omega_i))$$

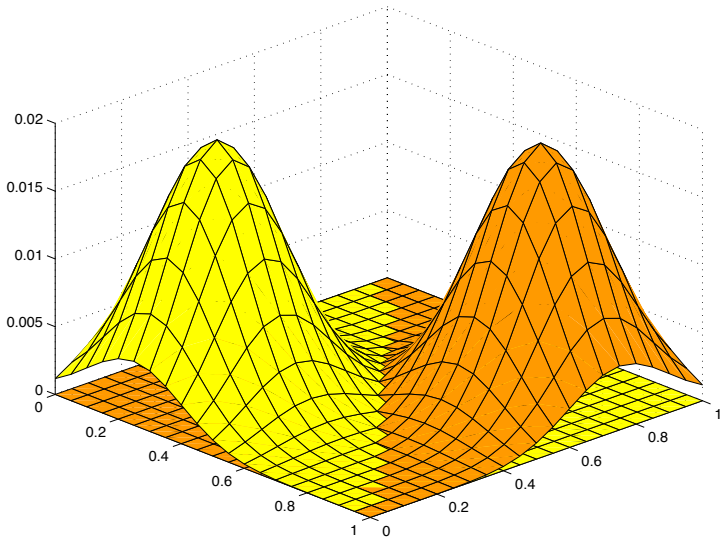
Cas 3 : matrices de cov. disctinctes

Les surfaces de décision obtenues sont des hyperquadratiques :

- paire d'hyperplans
- hypershères
- hyperellipsoïdes
- hyperhyperboloïdes

exemples de frontières

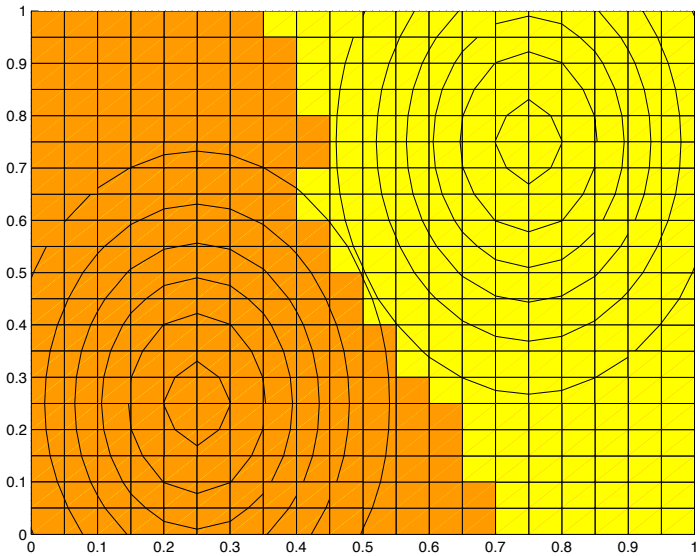




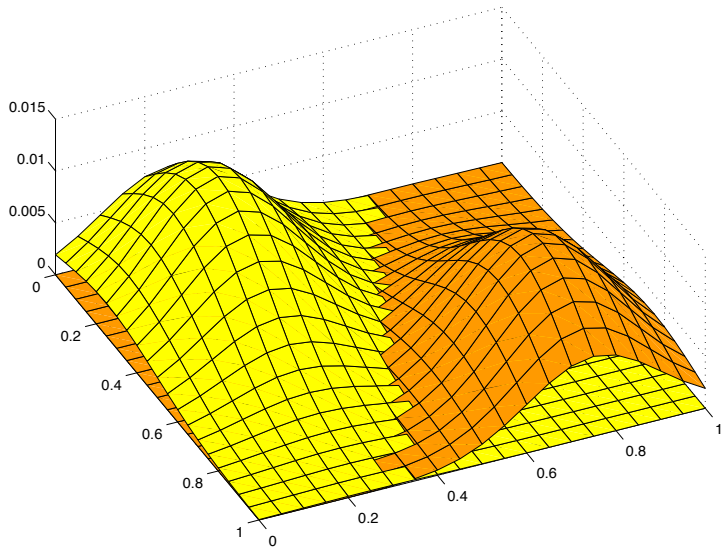
Cas1 :

$$\Sigma_{\omega_i} = \Sigma_{\omega_j} = \sigma^2 \mathrm{I}$$

exemples de frontières

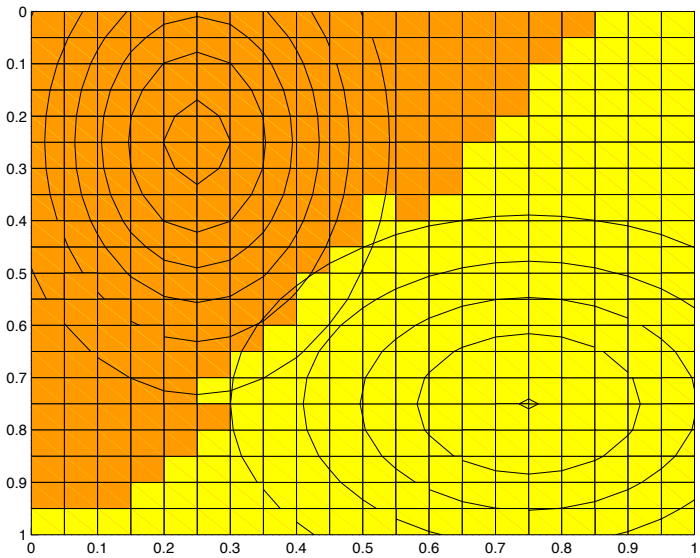


$\Sigma_{\omega_i} = \Sigma_{\omega_j}$

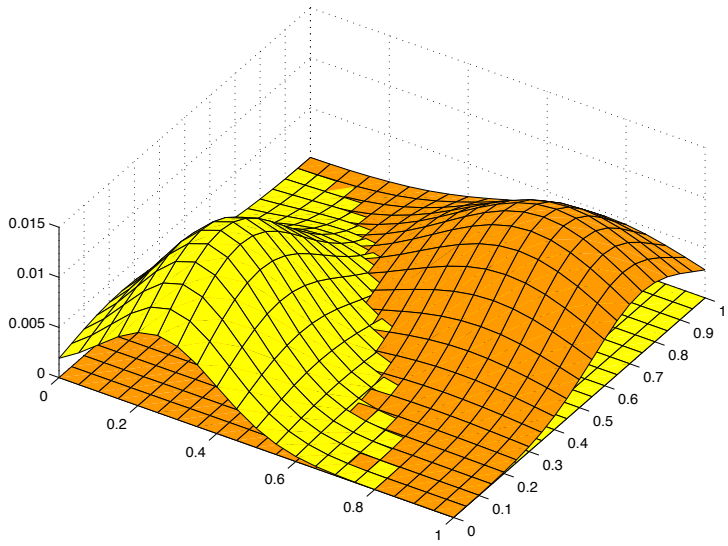


Cas2 :

exemples de frontières

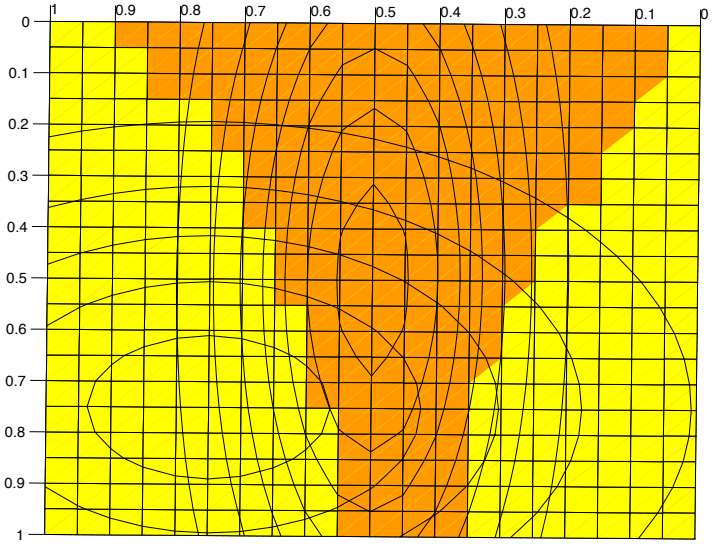


Quelconque

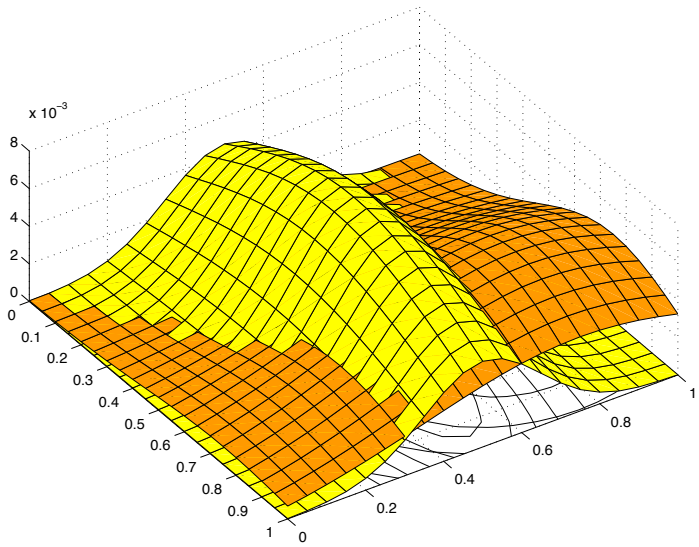


Cas3 :

exemples de frontières



Quelconque



Cas3 :

Remarques

- Hypothèse restrictive : toutes les classes suivent une loi gaussienne
- Ce n'est pas toujours (souvent) le cas.
- Il est alors possible de découper la classe en 2 deux sous-classes qui correspondent mieux à des lois normales.

Exercice : Classification Bayesienne

1. Soient deux classes suivant des lois normales et de probabilités a priori $P(C1) = P(C2) = 0.5$. On supposera que les formes sont des vecteurs de $\mathcal{R}^2$ et que :

$$\mu_{C1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \mu_{C2} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \ ; \ \Sigma_{C1} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \ \Sigma_{C2} = \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}$$

   Calculer la distance de Mahalanobis entre la forme $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ et chaque classe

   Pour info :

$$\Sigma_{C1}^{-1} = \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}, \ \Sigma_{C2}^{-1} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

2. Donner la forme de l'equation de la frontière de décision bayésienne entre les deux classes