SI221 Bases de l'apprentissage/Machine Learning for Pattern Recognition

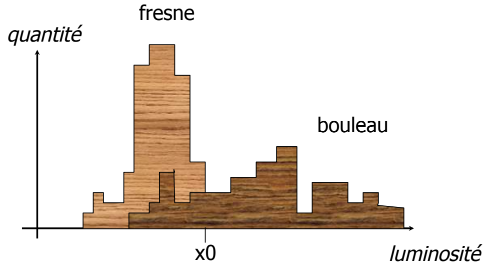
Décision bayésienne

Septembre 2018

Laurence Likforman-Sulem
Telecom ParisTech/IDS Image, Données, SIgnal
bureau : E 504
likforman@telecom-paristech.fr

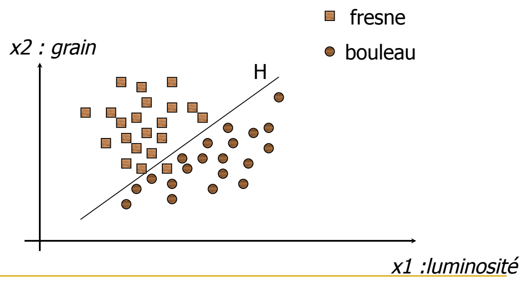
Un exemple

histogramme pour 2 classes



espace de représentation

2 paramètres : x1, x2



théorie de la Décision

- approche par modélisation
- minimise globalement le risque d'erreur
  - connaissances sur les distributions sous jacentes

Espace probabilisable: Ω , F(Ω)

- Ω : espace des formes (épreuves)
  Ω = {événements élémentaires}

- F(Ω) : ensemble de parties de Ω
  événement = sous-ensemble de Ω (partie de Ω)

Ω est divisé en ensembles disjoints (les classes)
ω1, ω2, ..., ωk avec

$$\omega_i \cap \omega_j = \phi \qquad \forall i \neq j$$

$$\bigcup_{i=1,K} \omega_i = \Omega$$

ω1, ω2, ..., ωk = système complet d'événements

Espace probabilisé: Ω , F(Ω), P

- Ω, F(Ω) : est muni d'une probabilité P
- P : F(Ω) --> [0,1]

P(Ω)= 1 P(ϕ) =0

$\forall A \;\; 0 \leq P(A) \leq 1$ $A \in$ F(Ω): A événement, partie de Ω

$\forall A, B \;\;$ si $\mathrm{A} \cap \mathrm{B} = \varnothing$ alors $\mathrm{P(A \cup B) = P(A) + P(B)}$

$A, B \in$ F(Ω)xF(Ω)

Espace des classes, ensemble de caractéristiques

- espace des classes -> ensemble des étiquettes

  J = {1, 2, ... K}

- espace de représentation -> ensemble des caractéristiques

  ex : R= R^d

variables/vecteurs aléatoires

- A chaque forme ω sont associées :
  - variable aléatoire de classe : C (ω)

    C : Ω--> J

  - vecteur aléatoire correspondant à la réalisation x : X(ω)

    X: $\Omega \longrightarrow R^d$

    $\omega \longmapsto x$

- On observe la réalisation x de X, issue d'une forme de classe C

Fonction de décision

- on cherche la fonction de décision :

  d : R^d --> J

- qui estime au mieux la classe , i.e.

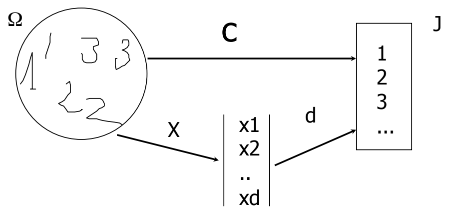
d(x)=estimation de la classe C(ω) de la forme ω dont la réalisation est X(ω)=x

exemple : chiffres

C : Ω--> J l'opérateur idéal (v.a de classe)

X : Ω--> R^d v.a associé à la représentation

d : R^d --> J fonction de décision



exemple : les planches de bois

Ω = ensemble des planches de frêne et de bouleau

J = { 1 = frêne , 2=bouleau }

X= (indice de luminosité, indice relatif à la texture)

ensemble des caractéristiques : R^2

on a $\omega 1 = C^{-1}$ (1) $\omega 2 = C^{-1}$ (2)

Loi de probabilité associée à un vecteur aléatoire

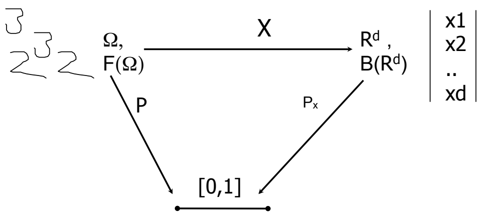
- vecteur aléatoire X: $\Omega \longrightarrow R^d$
- on probabilise R^d par P_x (loi image)
  - (R^d , B(R^d), P_x): espace probabilisé
  - B(R^d): tribu des Boréliens (famille d'intervalles dans R, de pavés dans R^d)
  - P_x: B(R^d) $\longrightarrow [0,1]$

$$P_X(A) = P(X^{-1}(A)) = P(\{\omega \in F(\Omega) | X(\omega) \in A\})$$

$A \in$ B(R^d)

noté $P(X \in A)$

Probabilité et loi de probabilité (loi image)



fonction de répartition, densité de probabilité

- $F_X : R^d \longrightarrow [0,1]$

$X=(X_1,X_2,\ldots,X_d)$

$x=(x_1,x_2,\ldots x_d)$

$F_X(x)=P_X(]-\infty,x_1]\times]-\infty,x_2]\times\ldots]-\infty,x_d])=P_X(X_1<x_1,X_2<x_2,\ldots X_d<x_d)$

noté $P(X<x)$

- densité de probabilité p_X associée à X (X admet une densité)

$$F_X(x)=\int_{]-\infty,x_1]\times\ldots]-\infty,x_d]} p_X(u)du \quad \text{noté} \quad \int_{]-\infty,x_1]\times\ldots]-\infty,x_d]} p\,(u)du$$

$$P_X(A)=P(X\in A)=\int_A p\,(x)dx \qquad A\in B(R^d)$$

densité de probabilité conditionnelle

- densités conditionnelles:
  - p (x | ωi) i=1,....K

$$P_{X|\omega i}(A)=P(X\in A|\omega i)=\int_A p(x|\omega i)dx \qquad A\in B(R^d)$$

$$P(X\in A|C=i)\Leftrightarrow P(\{\omega|X(\omega)\in A\,sachant\,C(\omega)=i\})$$

densités de probabilités

- densités jointes

$$P_{XY}(A,B) = P(X \in A \; et \; Y \in B) = \iint_{A,B} p(x,y)dxdy$$

$$P(X \in A, \omega_i) = \int_A p(x,\omega_i)dx = \int_A p(x|\omega_i)P(\omega_i)dx$$

- densités marginales

$$p(x) = \int_{R^d} p(x,y)dy \qquad p(y) = \int_{R^d} p(x,y)dx$$

Densités de probabilités (suite)

$$\int_{R^d} p(x|\omega_i)dx = 1 \text{ et } \int_{R^d} p(x)dx = 1$$

- densité mélange

$$p(x) = \sum_{i=1}^{K} p(x \mid \omega_i)\, P(\omega_i)$$

probabilités a priori, a posteriori

- probabilité a priori de la classe i : P(ωi)=P(C=i)

$$\sum_{i=1}^{K} P(\omega_i) = 1$$

- probabilité a posteriori

$$P(\omega_i|x) \Leftrightarrow P(C=i|X=x) \Leftrightarrow P(\{\omega|C(\omega)=i \; sachant \; X(\omega)=x\})$$

$$\sum_{i=1}^{K} P(\omega_i|x) = 1$$

formule de Bayes

$$P(\omega_i|x) = \frac{p(x|\omega_i)\,P(\omega_i)}{p(x)}$$

$P(\omega_i|x)$: probabilité a posteriori

$P(\omega_i)$: probabilité a priori

$p(x)$: densité "mélange"

P(A|B)=P(A,B)/P(B)

P(A,B)=P(B|A)P(A)

coûts de décision

- $\lambda(\omega i \mid \omega j)$= coût de décider d(x)=i pour une réalisation x appartenant à la classe ω j
- «classe» de rejet : $\omega 0$
  K+1 décisions , K classes
  $\lambda(\omega 0 \mid \omega j)$ = coût de rejet j=1,2,…K
- On cherche d : la règle de décision la moins coûteuse possible

risques (1)

- risque conditionnel de prendre la décision d(x)=i

  espérance de la fonction λ :

  $$C(d(x)=i \mid x) = C_i(x) = \sum_{j=1}^{K} \lambda(\omega i \mid \omega j)\, P(\omega j \mid x)$$
- risque conditionnel associé à la décision d(x)

  $$C(d(x) \mid x) = \sum_{j=1}^{K} \lambda(d(x) \mid \omega j)\, P(\omega j \mid x)$$

risques (2)

- risque associé à la règle de décision d
  - espérance du risque conditionnel

$$C(d) = \int_{R^d} C(d(x)|x)\, p(x)dx$$

Règle optimale d*

- d* est la règle qui minimise C(d)

-> Pour chaque x, on minimise C(d(x)|x)

- soit x donné, on décide d*(x)=i si

$$C_i(x) = C(\omega_i|x) \leq C_j(x) = C(\omega_j|x) \quad j = 1,2,....K \text{ sans rejet}$$
$$j = 0,1,2,....K \quad \text{avec rejet}$$

risque Bayésien

- risque bayésien = risque moyen associé à la règle de décision d* :

espérance du risque conditionnel C(d*(x)| x):

C* = E[C (d*(x)| x)] = $\int_{R^d} C(d^*(x)|x)\, p(x)dx$

- les réalisations sont classées par la règle bayésienne du risque conditionnel minimum :

$d^*(x) = Arg \min_{i=0,1,..K} C(\omega_i|x)$ et C* (x)= C(d*(x)| x) = $\min_{i=0,1,..K} C(\omega_i|x)$

$$C^* = \int_{R^d} C^*(x)p(x)dx$$

règle de Bayes à pénalisation symétrique

λ(ω0 | ωj)= λr ∀ j=1,..., K
λ(ωi | ωj)=1 si i ≠ j ∀ j=1,2,...K
λ(ωi | ωi)=0 ∀ i=1,..., K

Coût de rejeter x : $C_0(x) = C(\omega_0|x) = \sum_{j=1}^{K} \lambda_R P(\omega_j|x) = \lambda_R$

Coût de décider ωi :

$$C_i(x) = C(\omega_i|x) = \sum_{j=1, j\neq i}^{K} P(\omega_j|x) = 1 - P(\omega_i|x) = P(C \neq i|x)$$

décision

on décide d*(x)=i si

$C_i(x) = \sum_{j=1, j\neq i}^{K} P(\omega_j|x) = 1 - P(\omega_i|x) \leq C_j(x) \;\; \forall j = 1,2,..K$

et si $C_i(x) \leq C_0(x)$

On rejette si $C_0(x) \leq C_j(x) \quad \forall j = 1,2,..K$

Autre formulation : d*(x)=i si P (ωi | x) >= P (ωj | x) ∀ j=1,...,K
et $P(\omega_i|x) \geq 1 - \lambda_R$

Rejet si $\max_{j=1,..K} P(\omega_j|x) \prec 1 - \lambda_R$

probabilités d'erreur conditionnelles

- e*(x)=probabilité conditionnelle d'erreur associée à la décision optimale d*

$$e^*(x) = P(erreur|x) = P(C \neq d^*(x)|x) = 1 - P(d^*(x)|x)$$

$$e^*(x) = 1 - \underset{i}{Max}\, P(\omega i|x)$$

- $e_i(x)$ =probabilité conditionnelle d'erreur associée à la décision d(x)=i (optimale ou non)

$$e_i(x) = P(C \neq i|x) = 1 - P(C = i|x) = 1 - P(\omega i|x)$$

influence de λ_R

Seuil de rejet : $0 \leq \lambda_R \leq \frac{K-1}{K}$

Espace de représentation : $R = R_0 \cup R_A$

$$R_A = \bigcup_{i=1,\dots K} R_i$$

Influence du seuil de rejet :

$$\lambda_1 \leq \lambda_2 \text{ alors } R_A(\lambda_1) \subset R_A(\lambda_2)$$

probabilités acceptation/rejet

Probabilité de classement dans la classe ωi $P_{Ai} = \int_{Ri} p(x)dx$

Probabilité globale d'acceptation $P_A = \int_{RA} p(x)dx = \sum_{i=1}^{K} P_{Ai}$

Probabilité de rejet $P_R = \int_{R0} p(x)dx$

probabilités d'erreur

Probabilité d'erreur de classement dans ωi

$$P_{Ei} = \int_{Ri} e^*(x)p(x)dx$$

$$\text{avec } e^*(x) = P(erreur|x)$$

Probabilité globale d'erreur : $P_E = \int_{R_A} e^*(x)p(x)dx$

probabilité de bon classement

Dans la classe ωi :

$$P_{C_i} = \int_{R_i} (1 - e^*(x))\, p(x)dx$$

Probabilité globale de bon classement

$$P_C = \int_{RA} (1 - e^*(x))\, p(x)dx = \sum_{i=1}^{K} Pc_i$$

Suite…

On a :

$$P_A + P_R = 1$$
$$P_A = P_C + P_E$$
$$P_C + P_E + P_R = 1$$

P_A, P_E, P_R et P_C dépendent de λ_R

risque Bayésien

$$C^* = \int_R C^*(x)\, p(x)dx$$ Avec C*(x)=C(d*(x)|x)

$$C^* = \int_{R_A} C^*(x)\, p(x)dx + \int_{R_0} C^*(x)\, p(x)dx$$

Règle des coûts 0-1 :

$$C^*(x) = e^*(x) = 1 - \max_{i=1,..K} P(\omega i|x) \quad pour \;\; x \in R_A$$

Règle des coûts 0-1 : $C^*(x) = e^*(x)$

$$C^* = \int_{R_A} e^*(x)p(x)dx + \int_{R_0} \lambda_R p(x)dx$$

$C^*=P_E + \lambda_R P_R$

C^* = risque minimum=risque bayésien

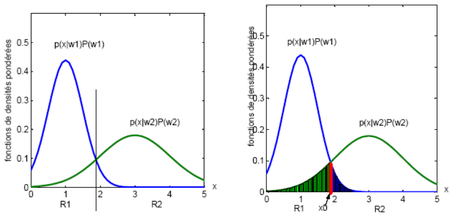
rapport de vraisemblance

d*(x)=1 si

$$\Lambda(x) = \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \frac{\lambda(\omega_1|\omega_2)P(\omega_2)}{\lambda(\omega_2|\omega_1)P(\omega_1)} = \eta$$

d*(x)=2 sinon

probabilité d'erreur



$$P(erreur) = P(X \in R_1, \omega_2) + P(X \in R_2, \omega_1)$$
$$= P(X \in R_1|\omega_2)P(\omega_2) + P(X \in R_2|\omega_1)P(\omega_1)$$
$$= \int_{R1} p(x|\omega_2)P(\omega_2)dx + \int_{R2} p(x|\omega_1)P(\omega_1)dx$$

Exemple : vérification d'identité

- problème à 2 classes
- ω1 : clients du système
- ω2 : imposteurs
- Décision :

$$s(x) = \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \eta$$

erreurs de type I et II

H0 - hypothèse nulle: absence d'imposteur

- P_E^1 : Erreur de type 1 (rejeter H0 de manière erronée): rejeter un client ou faux rejet (fausse alarme, faux positif)
- P_E^2 : Erreur de type 2 : (accepter H0 de manière erronée): accepter un imposteur ou fausse acceptation (détection manquée, faux négatif)

	ω1: client	ω2:imposteur
d(x)=1	correct	type 2
d(x)=2	type 1	correct

Suite…

$$P_E^2 = P(\text{décider } \omega 1, \omega 2 \text{ est vrai}) \quad \text{prob. jointe}$$

$$P_E^2 = P(d(x) = 1 | \omega 2) P(\omega 2) = \int_{R_1} p(x | \omega 2) P(\omega 2) dx$$

$$P_E^2 = \int_{R_1} P(\omega 2 | x) p(x) dx$$

$$P_E^1 = P(\text{décider } \omega 2, \omega 1 \text{ est vrai})$$

$$P_E^1 = \int_{R_2} p(x | \omega 1) P(\omega 1) dx = \int_{R_2} P(\omega 1 | x) p(x) dx$$

$$P_E = P_E^1 + P_E^2$$

vérification d'identité : FRR-FAR

- espace de représentation
- probabilités conditionnelles

Taux de faux rejet = FRR= P(d(x)=2|ω1 est vrai)
Taux de fausse acceptation = FAR= P(d(x)=1|ω2 est vrai)

$$FRR = \int_{R2} p(x|\omega_1)dx$$

$$FAR = \int_{R1} p(x|\omega_2)dx$$

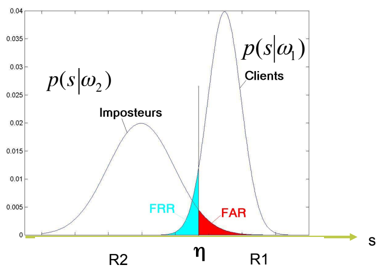
$$P_E = P(\omega_1)FRR + P(\omega_2)FAR$$

FAR-FRR : espace des scores

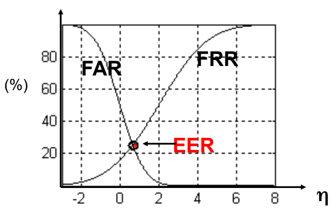
$$s(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \eta$$

$$FRR = \int_{R2} p(s|\omega_1)ds$$

$$FAR = \int_{R1} p(s|\omega_2)ds$$



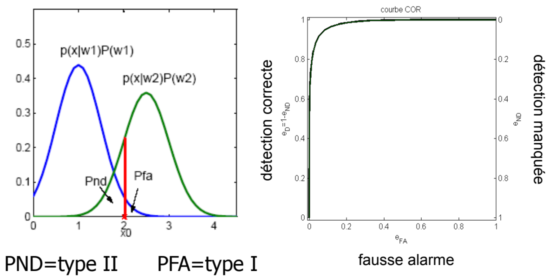
EER : Equal Error rate



Courbe ROC

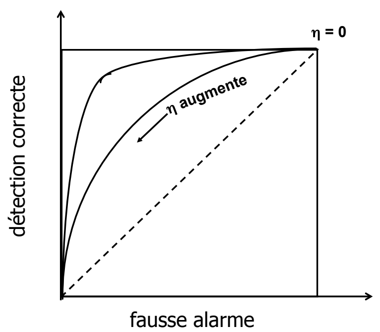
- COR : Caractéristique Opérationnelle du Récepteur
  - ω1 : absence d'objet
  - ω2 : présence d'objet
- incrémentation du seuil de décision
  - courbe Probabilité détection correcte en fonction probabilité de fausse alarme

$$si\ \ x > x_0\ \ (seuil)\ \text{alors présence de cible}\ (\omega 2)$$



PND=type II PFA=type I

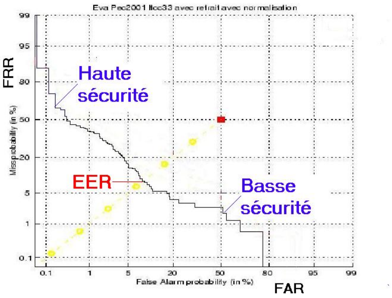
Courbe ROC



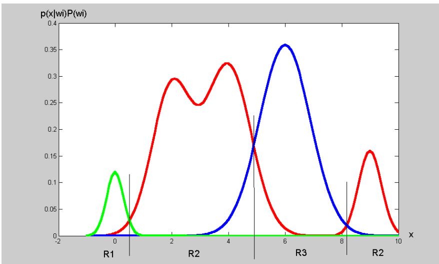
aire sous la courbe ROC

Courbe DET

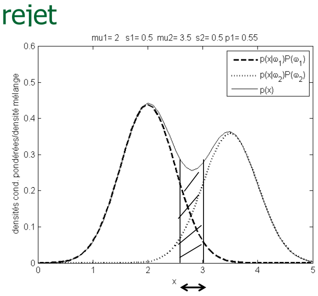
[chollet]



Zones de décision



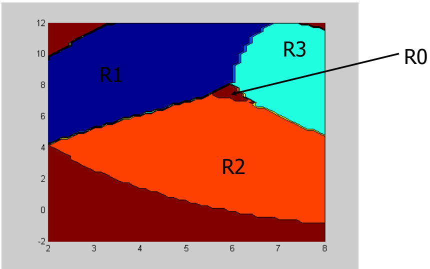
zone de rejet



R1 R0 R2

$$P_{rejet} = \int_{R0} p(x)dx$$

zone de rejet



distributions normales

- M : vecteur moyenne classe ω
  - E[X|ω]

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_d \end{pmatrix} \qquad M = \begin{pmatrix} E(X_1) \\ E(X_2) \\ E(X_d) \end{pmatrix}$$

- Σ: matrice de covariance classe ω
  - E[(X-M)(X-M)^t|ω]
  - Σ est symétrique définie positive

$$p\,(x|\omega) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}}\, e^{-\frac{1}{2}(x-M)^t\Sigma^{-1}(x-M)}$$

distributions normales (suite)

- Σ: matrice de covariance
  - variance σ^2_i (caractéristique Xi)
  - covariance σ_{ij}: (caractéristiques Xi et Xj)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

$$\sigma_i^2 = E\left[(X_i - E(X_i))(X_i - E(X_i))\right]$$
$$\sigma_{ij} = E\left[(X_i - E(X_i))(X_j - E(X_j))\right]$$
$$= E(X_iX_j) - E(X_i)E(X_j)$$

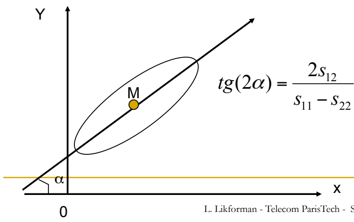
ellipsoide de Mahalanobis

- lieu des points

$$U = \begin{bmatrix} X \\ Y \end{bmatrix}$$

$$(U-M)^t \Sigma^{-1} (U-M) = Cte$$

- orientation



$$tg(2\alpha) = \frac{2s_{12}}{s_{11} - s_{22}} \quad avec\ \Sigma = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$$

ellipsoide de Mahalanobis

- décomposition: Σ symétrique

$$\Sigma = V * \Sigma_d * V^t$$

$$\text{V : matrice des vecteurs propres} \quad V = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

$$\Sigma_d = \begin{bmatrix} \lambda 1 & 0 \\ 0 & \lambda 2 \end{bmatrix} \quad \text{matrice diagonalisee}$$

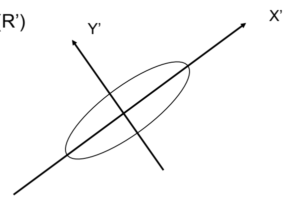
$$\Sigma = R^{-1} * \Sigma_d * R$$

$$\text{R : matrice de passage de (R) à (R')} \quad R = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

ellipsoide de Mahalanobis

- grand axe, petit axe
  - repère de l'ellipsoide (R')

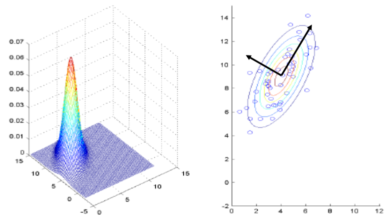
$$\frac{X'^2}{\lambda_1} + \frac{Y'^2}{\lambda_2} = Cte$$



$$1/2\ \ grand\ \ axe : \propto \sqrt{\max(\lambda_1, \lambda_2)}$$

$$1/2\ \ petit\ \ axe : \propto \sqrt{\min(\lambda_1, \lambda_2)}$$

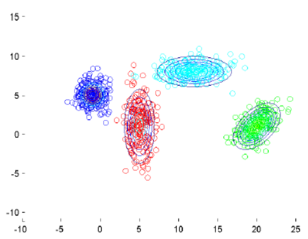
distributions normales (suite)



Conclusion : décision Bayésienne

- cadre d'application
  - Les distributions p(x|wi), P(wi) doivent etre connues
  - Sinon, il faut pouvoir les estimer (#échantillons suffisant)
- avantages/inconvénients
  - Classification optimale
  - Mise en œuvre simple
  - Estimation peut être délicate
- complexité
  - paramètres des modèles de classe

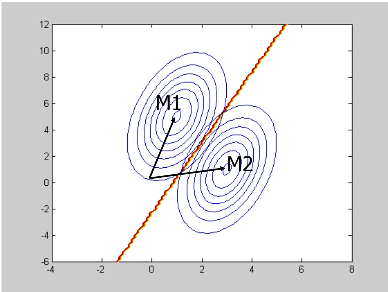
ellipsoide de Mahalanobis



2 lois gaussiennes : frontière

$\Sigma_i = \Sigma$

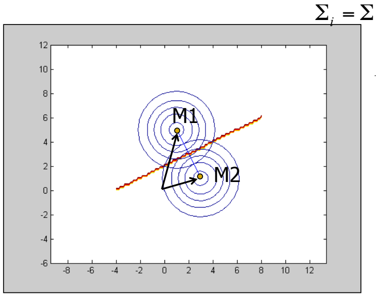
$P(\omega_1) = P(\omega_2)$



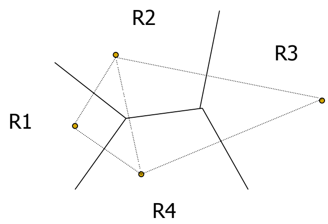
2 lois gaussiennes : frontière

$\Sigma_i = \Sigma \; diagonale : \propto \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

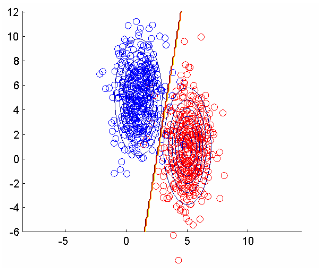
$P(\omega_1) = P(\omega_2)$



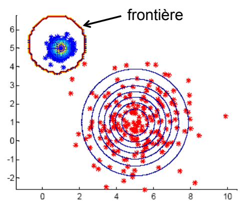
4 classes (mêmes matrices de covariance)

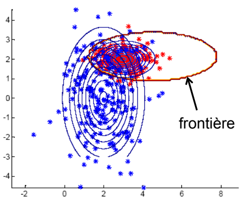


Frontières : hyperplans ou hyperquadriques



cercle, ellipse





Conclusion : décision Bayésienne

- cadre d'application
  - Les distributions p(x|wi), P(wi) doivent etre connues
  - Sinon, il faut pouvoir les estimer (#échantillons suffisant)
- avantages/inconvénients
  - Classification optimale
  - Mise en œuvre simple
  - Estimation peut être délicate
- complexité
  - paramètres des modèles de classe