

# Introduction à l'Apprentissage Automatique

Cours 4 - apprentissage supervisé : approches statistiques,  
classifieur bayésien naïf, modèle Gaussien

Thomas Pellegrini, équipe SAMoVA, IRIT,  
[thomas.pellegrini@irit.fr](mailto:thomas.pellegrini@irit.fr)

IRIT - UPS

# Approche bayésienne

- ▶ Approche bayésienne : méthode statistique à apprentissage supervisé
- ▶ On fait des hypothèses sur la distribution statistique des points de chaque classe
- ▶ Estimation paramétrique : paramètres d'une loi hypothèse, par ex. loi Gaussienne,
- ▶ Estimation non-paramétrique : k plus proches voisins, fenêtres de Parzen

## Principe de la classification bayésienne

- ▶ Distributions : modèles des classes *1 distribution par classe*
- ▶ Utilisées pour prédire la classe la plus probable d'une observation *"vraisemblance"* *"likelyhood"*
- ▶ Garantit un taux global d'erreur minimal *minimisation de la vraisemblance*

# Décision à partir des probabilités a priori

Exemple 1 : deux classes  $c_1$  et  $c_2$

- ▶  $P(c_1) = 0.8$  et  $P(c_2) = 0.2$
- ▶ Classifieur rudimentaire : prédit toujours la classe la plus probable  $c_1$
- ▶ Que vaut  $P_{\text{erreur}}$ ?  $P_{\text{erreur}} = \frac{P(c_2|c_1) P(c_1) + P(c_1|c_2) P(c_2)}{P(c_2)}$   
$$P_{\text{erreur}} = \frac{0 \times 0,8 + 1 \times 0,2}{0,2} = 0,2$$

Exemple 2

- ▶ Classifieur un peu moins rudimentaire : prédit aléatoirement à 80% la classe  $c_1$  et à 20% la classe  $c_2$
- ▶ Que vaut  $P_{\text{erreur}}$  ?

$$P_{\text{erreur}} = 0,2 \times 0,8 + 0,8 \times 0,2 = 0,32$$

# Approche bayésienne : principe

- ▶ Prédire la classe  $c_1$  si  $P(c_1|x) \geq P(c_2|x)$
- ▶ Prédire la classe  $c_2$  si  $P(c_2|x) > P(c_1|x)$

Difficulté : ces probabilités, dites *a posteriori*, ne sont pas directement accessibles

# Loi de Bayes

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)}$$

Diagram illustrating the components of the Bayes formula:

- prob a posteriori ( $c_i$ )**:  $P(c_i|x)$
- évidence**:  $P(x)$
- prob a priori**:  $P(c_i)$
- prob jointe**:  $P(x|c_i)P(c_i)$
- raisons d'assurance**:  $P(x|c_i)$

avec

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

évidence : loï des probas totales

Principe : après avoir vu un échantillon des données, en quoi cela modifie mes décisions a priori ?

$$P(c_i) \rightarrow P(c_i) \times P(x|c_i)$$

a priori

# Loi de Bayes : exemple à deux classes

- ▶ Deux classes,  $c_1$  : canard,  $c_2$  : poulet
- ▶ Observation  $X$ , V.A. discrète : la couleur des ailes

$$P(\text{canard} | \text{ailes grises}) = \frac{P(\text{ailes grises} | \text{canard}) P(\text{canard})}{P(\text{ailes grises})}$$

$P(\text{ailes grises}) = P(\text{ailes grises} | \text{canard}) P(\text{canard}) + P(\text{ailes grises} | \text{poulet}) P(\text{poulet})$

# Approche bayésienne : principe

- ▶ Prédire la classe  $c_1$  si  $\underline{P(c_1|x)} \geq \underline{P(c_2|x)}$
- ▶ Prédire la classe  $c_2$  sinon

( $\Leftrightarrow$ )

- ▶ Prédire la classe  $c_1$  si  
 $\cancel{P(x|c_1)P(c_1)/P(x)} \geq \cancel{P(x|c_2)P(c_2)/P(x)}$
- ▶ Prédire la classe  $c_2$  sinon

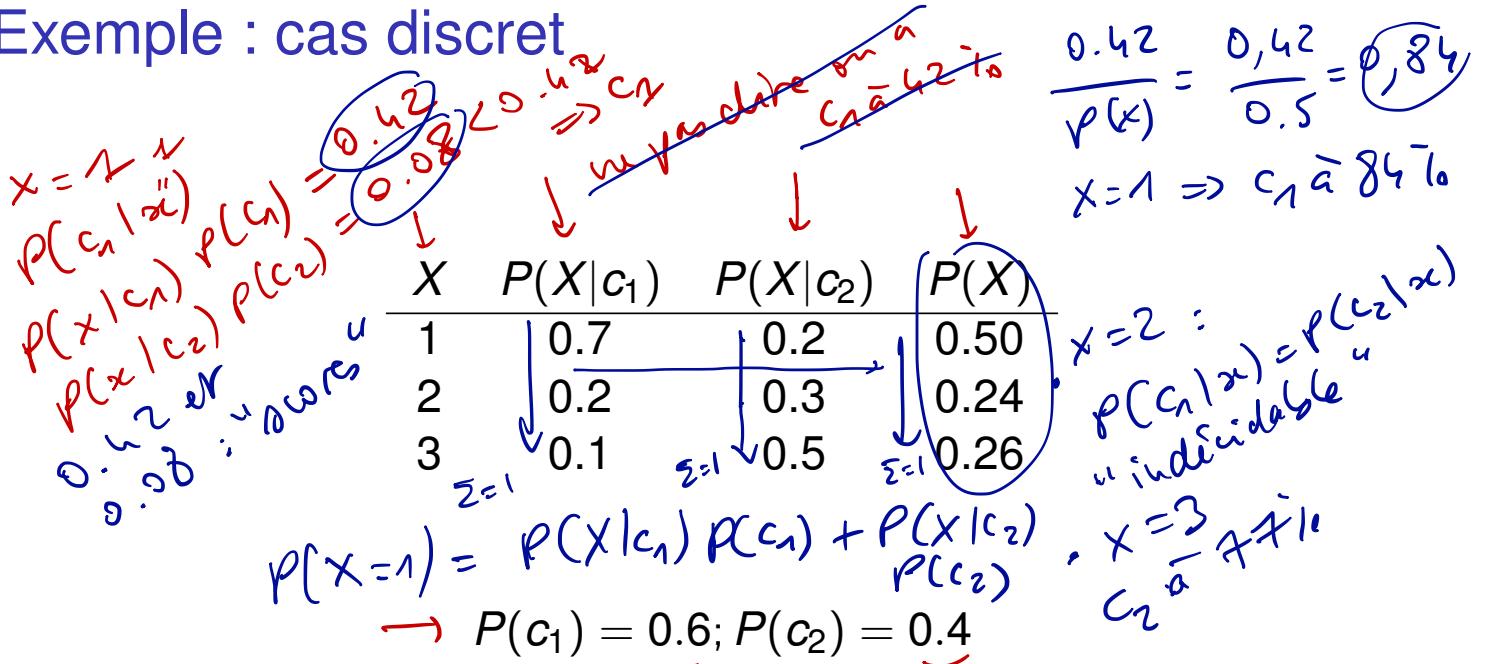
$P(x) = \text{facteur de normalisation}$

( $\Rightarrow$ )

- ▶ Prédire la classe  $c_1$  si  $\underline{P(x|c_1)P(c_1)} \geq \underline{P(x|c_2)P(c_2)}$
- ▶ Prédire la classe  $c_2$  sinon

Score ( $c_1$ ) =  $\frac{P(x|c_1) P(c_1)}{\text{facteur de normalisation}}$   
Ce n'est pas une proba mais une valeur réelle.

# Exemple : cas discret



Quelle classe choisir si on observe  $x = 1$ ,  $x = 2$  ou  $x = 3$  ?

# Cas continu

- ▶ La variable aléatoire  $X$  est continue et admet une densité de probabilité  $p$

$p :$  

rappel :  $P(X \in A) = \int_A p(x)dx$     et     $\int_{\mathbb{R}} p(x)dx = 1$  

# Cas continu

- ▶ La variable aléatoire  $X$  est continue et admet une densité de probabilité  $p$

rappel :  $P(X \in A) = \int_A p(x)dx$       et       $\int_{\mathbb{R}} p(x)dx = 1$

- ▶ Densité conditionnelle :

$$P(X \in A | c_i) = \int_A \underbrace{p(x|c_i)}_{\downarrow} dx \quad \text{et} \quad \int_{\mathbb{R}} p(x|c_i)dx = 1$$

exemple : Gaussiane

# Cas continu

- ▶ La variable aléatoire  $X$  est continue et admet une densité de probabilité  $p$

rappel :  $P(X \in A) = \int_A p(x)dx$       et       $\int_{\mathbb{R}} p(x)dx = 1$

- ▶ Densité conditionnelle :

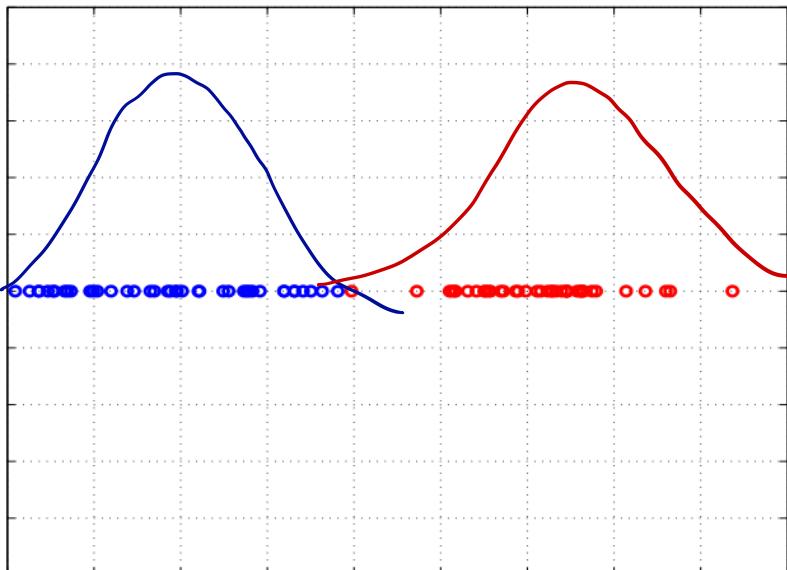
$$P(X \in A | c_i) = \int_A p(x|c_i)dx \quad \text{et} \quad \int_{\mathbb{R}} p(x|c_i)dx = 1$$

On a aussi, par la loi des probabilités totales :

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

# Exemple : cas Gaussien 1-d

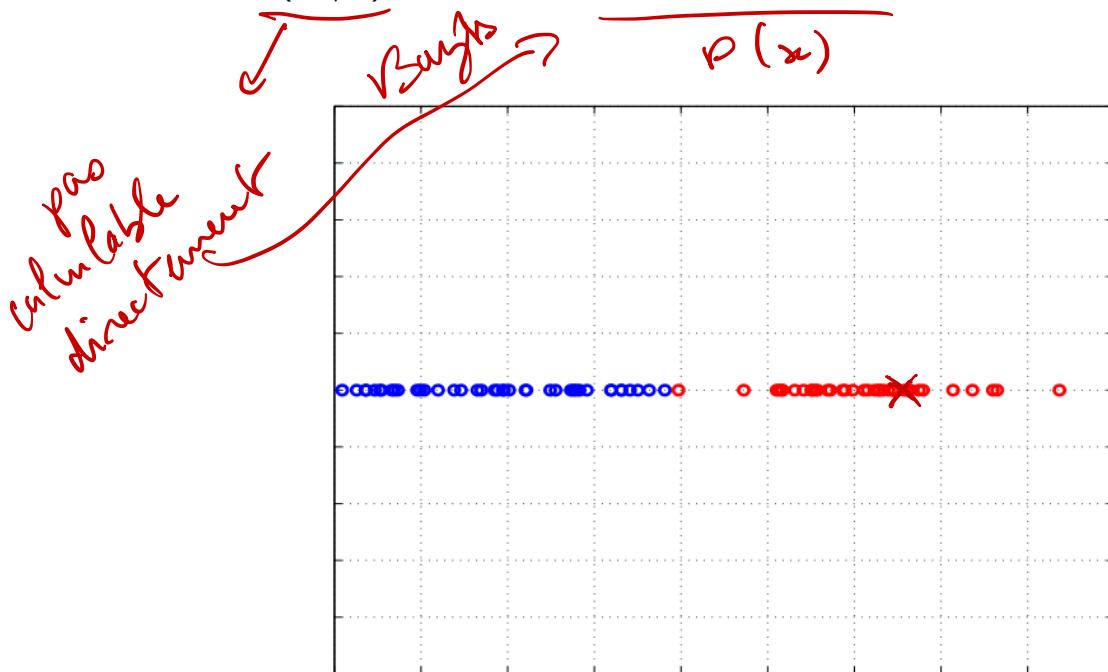
Deux classes :  $c_1$  et  $c_2$



# Exemple : cas Gaussien 1-d

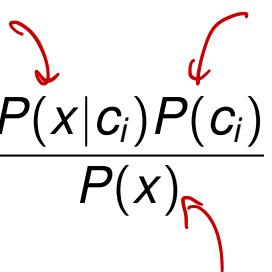
Soit une observation  $x$  : appartient-elle à  $c_1$  ou  $c_2$  ?

► i.e.  $P(c_i|x) = ??$   $\frac{P(x|c_i) \times P(c_i)}{P(x)}$



# Exemple : cas Gaussien 1-d

Comment faire ? On utilise l'approche bayésienne :

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)}$$


- ▶ On doit estimer chacun des trois termes de droite

# Exemple : cas Gaussien 1-d

## Probabilités *a priori*

Quelle répartition global des exemples entre les deux classes ?

- ▶  $P(c_1) \approx \boxed{N_1}/N$
- ▶  $P(c_2) \approx \boxed{N_2}/N$

# Exemple : cas Gaussien 1-d

## Probabilités *a priori*

Quelle répartition global des exemples entre les deux classes ?

- ▶  $P(c_1) \approx N_1/N$
  - ▶  $P(c_2) \approx N_2/N$
- )
- a priori

## Vraisemblances : $P(x|c_1)$ et $P(x|c_2)$ ?

- ▶ On doit connaître la distribution de  $c_i$
- ▶ Distribution uniforme, exponentielle, (de Poisson), normale (Gaussienne), etc. (Y a-t-il un intrus ?)

la distribution

# Exemple : cas Gaussien 1-d

## Probabilités *a priori*

Quelle répartition global des exemples entre les deux classes ?

- ▶  $P(c_1) \approx N_1/N$
- ▶  $P(c_2) \approx N_2/N$

## Vraisemblances : $P(x|c_1)$ et $P(x|c_2)$ ?

- ▶ On doit connaître la distribution de  $c_i$
- ▶ Distribution uniforme, exponentielle, de Poisson, normale (Gaussienne), etc. (Y a-t-il un intrus ?)

## Évidence ou facteur de normalisation : $P(x)$

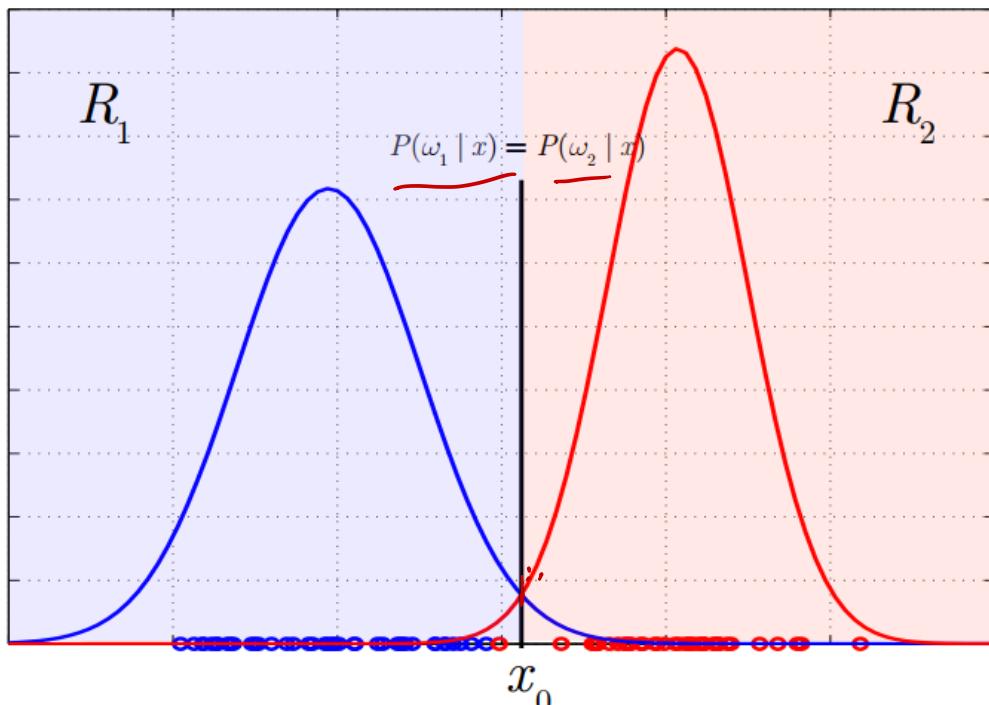
- ▶  $P(x) = \underbrace{p(x|c_1)}_{\text{Nécessaire ? Comparer}} P(c_1) + \underbrace{p(x|c_2)}_{\text{Nécessaire ? Comparer}} P(c_2)$

# Exemple : cas Gaussien 1-d

Frontière de décision

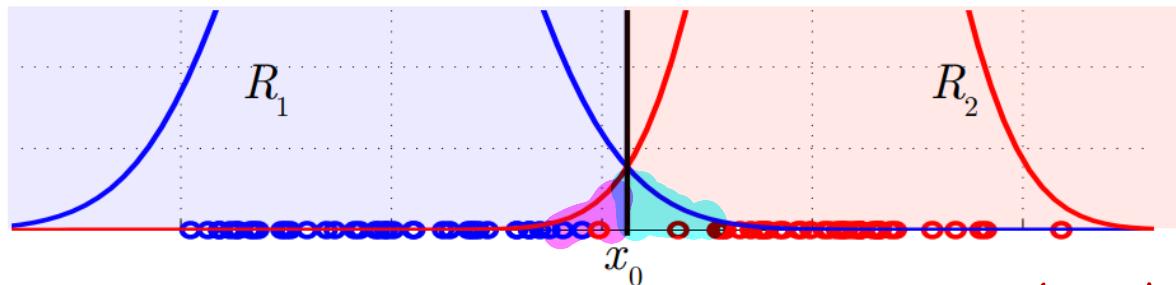
- ▶ Hypothèse loi normale :  $P(x|c_i) = \mathcal{N}(\mu_i, \sigma_i)$

vraisemblance



# Exemple : cas Gaussien 1-d

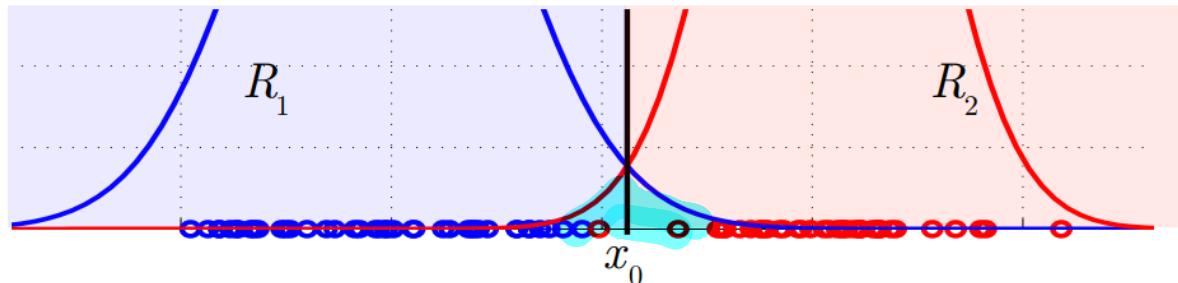
## Erreurs de classification



: erreurs : je prédis  $C_1$  à la place de  $C_2$

# Exemple : cas Gaussien 1-d

## Erreurs de classification

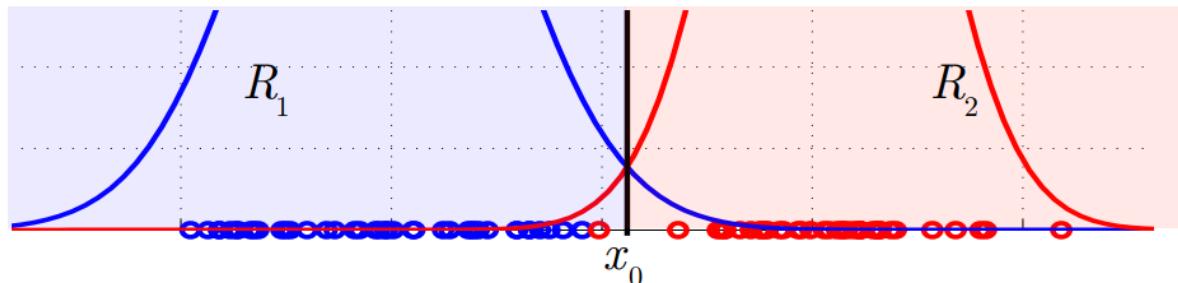


Probabilité globale d'erreur

$$P_{\text{erreur}} = \int_{R_1} p(x|c_2)P(c_2)dx + \int_{R_2} p(x|c_1)P(c_1)dx$$

# Exemple : cas Gaussien 1-d

## Erreurs de classification



## Probabilité globale d'erreur

$$P_{\text{erreur}} = \int_{R_1} p(x|c_2)P(c_2)dx + \int_{R_2} p(x|c_1)P(c_1)dx$$

La règle de décision Bayésienne garantit d'avoir  $P_{\text{erreur}}$  minimale

# Pénaliser plus certaines erreurs de classification

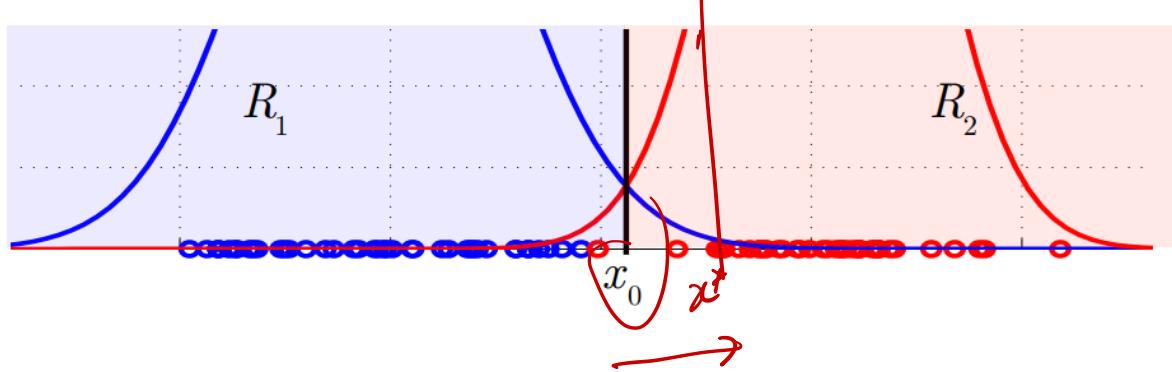
On peut définir une matrice de coûts :

- ▶  $\lambda(c_i|c_j) = \lambda_{ij}$  : coût de classer un objet de classe  $c_j$  dans la classe  $c_i$

# Pénaliser plus certaines erreurs de classification

On peut définir une matrice de coûts :

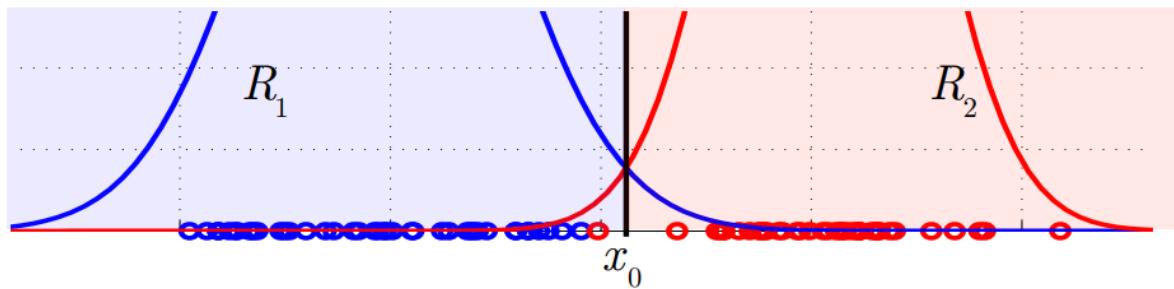
- ▶  $\lambda(c_i|c_j) = \lambda_{ij}$  : coût de classer un objet de classe  $c_j$  dans la classe  $c_i$
- Erreur minimale pour toutes les classes



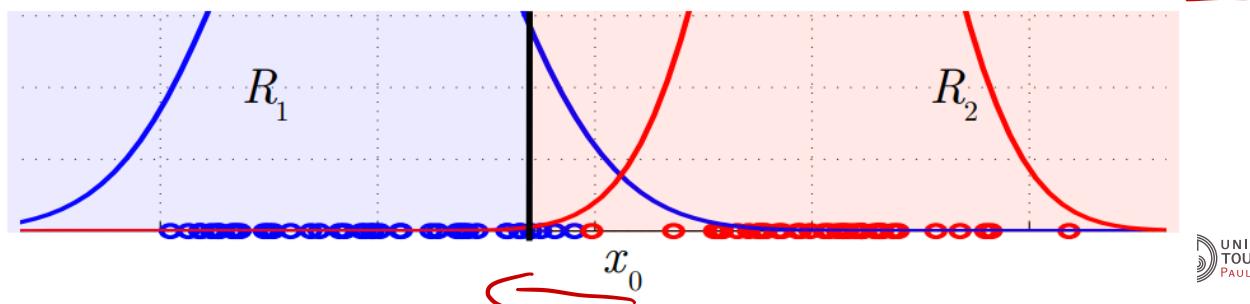
# Pénaliser plus certaines erreurs de classification

On peut définir une matrice de coûts :

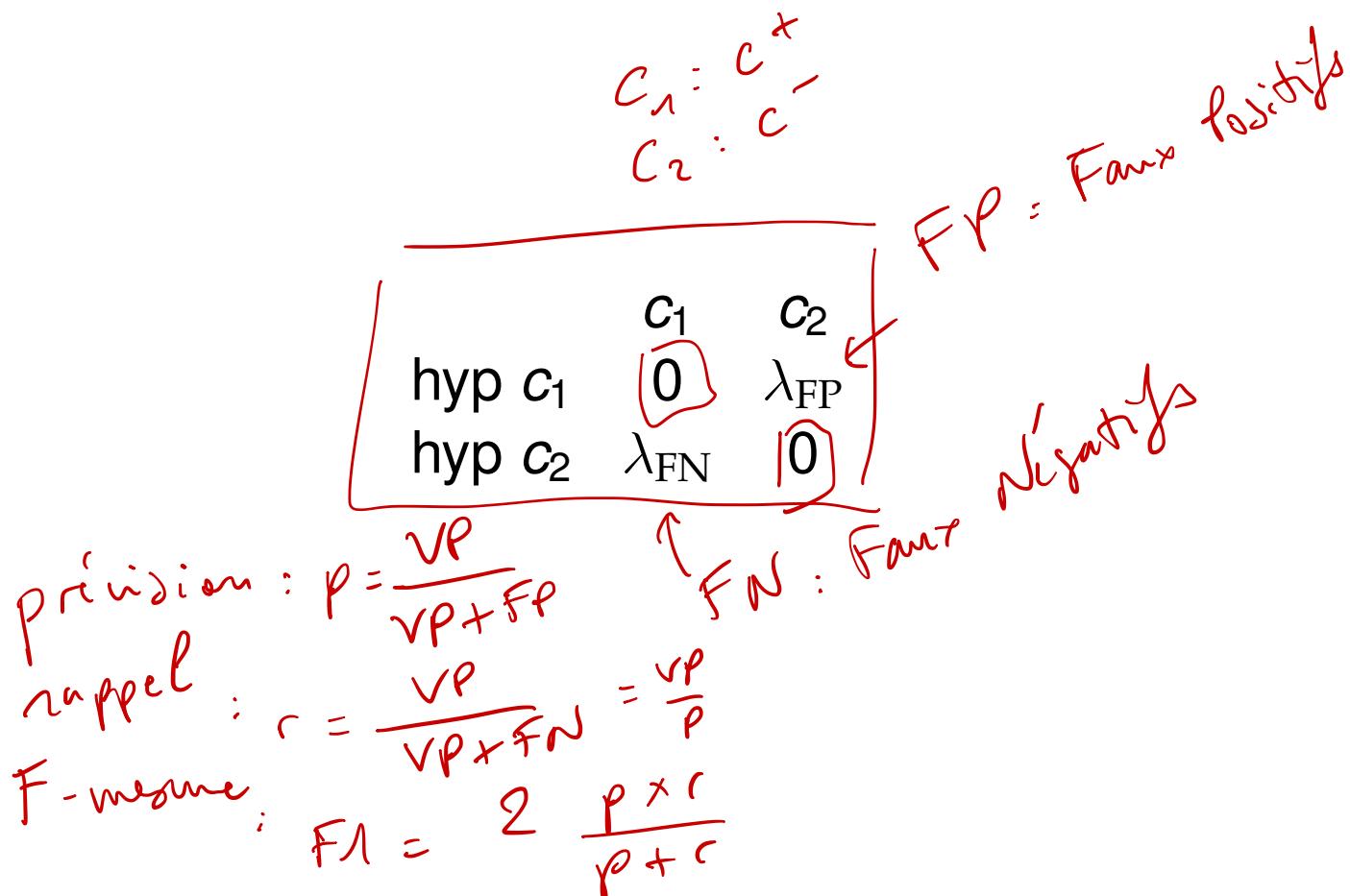
- ▶  $\lambda(c_i|c_j) = \lambda_{ij}$  : coût de classer un objet de classe  $c_j$  dans la classe  $c_i$
- Erreur minimale pour toutes les classes



- Par ex., si la classe  $c_2$  est plus importante : prendre  $\lambda_{21} < \lambda_{12}$



# Exemple à deux classes



# Option de rejet

Pour de nombreuses applications, prendre une mauvaise décision peut avoir un impact très important

- ▶ Ajouter une  $(K + 1)$ ème classe dite classe de rejet avec un coût  $\lambda$
- ▶ Nouvelle stratégie de décision :

$$\rightarrow h(x) = \begin{cases} K + 1 \\ \operatorname{argmax} P(c_j|x) \end{cases} \quad \begin{array}{l} \text{si } P(c_j|x) < 1 - \lambda, \forall j \in [1, K] \\ \text{sinon, avec } j \in [1, K] \end{array}$$

$\text{argmax } P \rightsquigarrow$  indice  
 $\text{mp. argmax } (P)$

# Risque de Bayes

- Risque conditionnel : espérance de la fonction discrète  $\lambda$  :

$$C_i(x) = C(c_i|x) = \underbrace{\sum_{j=1}^K}_{\text{---}} \lambda(c_i|c_j) P(c_j|x)$$

# Risque de Bayes

- ▶ *Risque conditionnel* : espérance de la fonction discrète  $\lambda$  :

$$C_i(x) = C(c_i|x) = \sum_{j=1}^K \lambda(c_i|c_j)P(c_j|x)$$

- ▶ *Risque de Bayes* : espérance de  $C(h(x)|x)$  où  $h^*(x)$  est la décision bayésienne pour l'observation  $x$  :

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

# Décision Bayésienne

- Minimise le *risque conditionnel* et par là-même le *risque de Bayes* :

$$h^*(x) = \operatorname{argmin} \sum_{j=1}^K \lambda(c_i|c_j) P(c_j|x)$$

$$h^*(x) = \operatorname{argmin} \sum_{j=1}^K \lambda(c_i|c_j) P(x|c_j) P(c_j)$$

# Critère MAP : Maximum A Posteriori

- ▶ Si les coûts sont les mêmes pour toutes les classes, la DB devient le critère MAP :

$$h^*(x) = \operatorname{argmax} \underbrace{P(x|c_j)}_{\text{---}} \underbrace{P(c_j)}_{\text{---}}$$

# Critère ML : Vraisemblance Maximale

- ▶ Si de plus, les classes sont équiprobables, la DB devient le critère ML :

ML, Maximum Likelihood

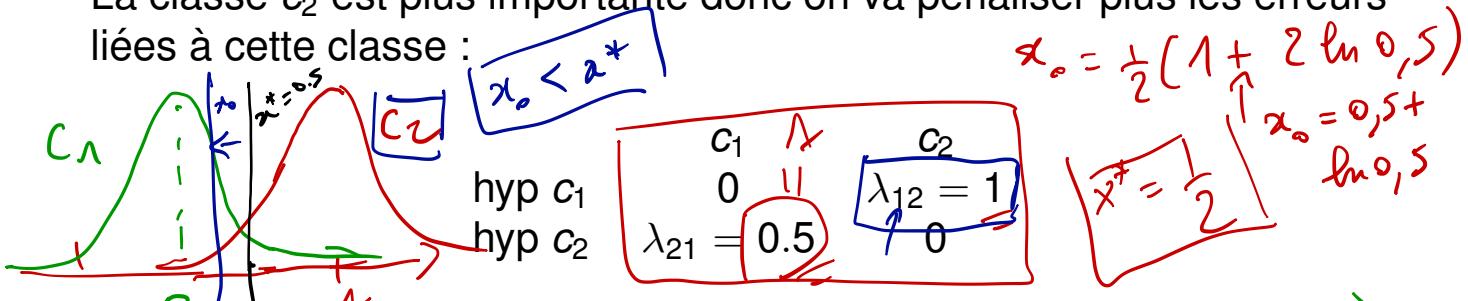
$$h^*(x) = \operatorname{argmax} P(x|c_j)$$

~~P(c<sub>j</sub>)~~

# Exemple : deux classes et Gaussien 1-d

$$p(x|c_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \text{ et } p(x|c_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-1)^2}{2}\right)$$

La classe  $c_2$  est plus importante donc on va pénaliser plus les erreurs liées à cette classe :



La règle du risque de Bayes moyen dit que l'on classe une observation  $x$  en  $c_1$  si :

$$\lambda_{12}P(c_1)p(x|c_1) > \lambda_{21}P(c_2)p(x|c_2)$$

$$-\frac{x_0^2}{2} = -\frac{(x-1)^2}{2} + \ln(0.5)$$

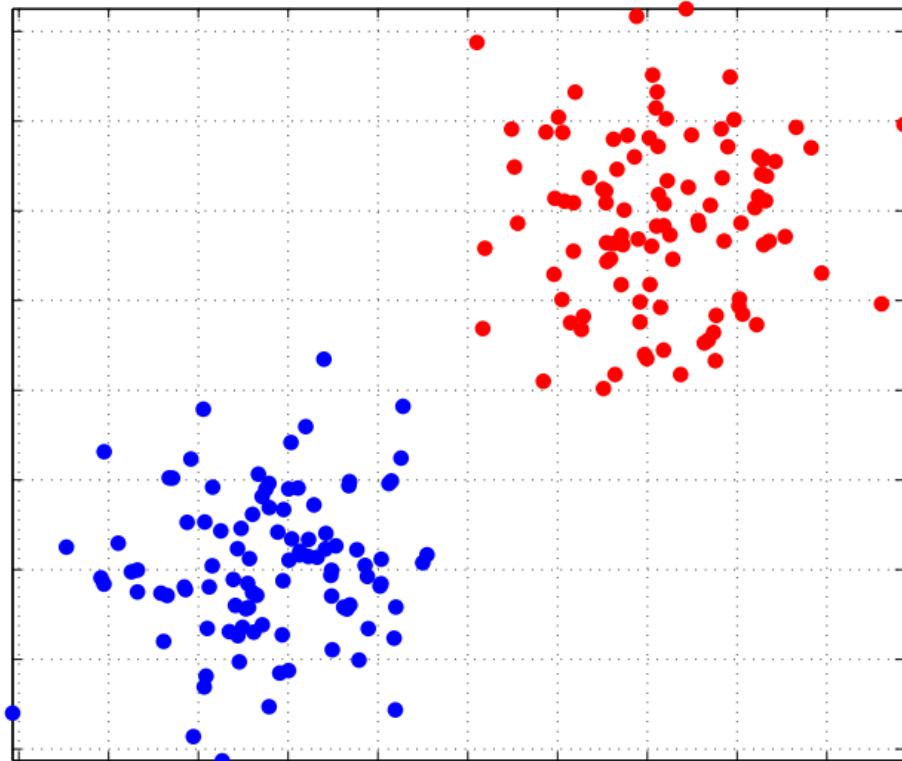
- Quel est le seuil de décision  $x_0$  ?

- Quel est le seuil de décision Bayésien  $x^*$  ?

$$e^{-\frac{x_0^2}{2}} = \frac{0.5}{e^{-\frac{(x-1)^2}{2}}} \quad \dots x_0 = \frac{1}{2}(1 + 2 \ln 0.5)$$

$$x_0 = 0.5 + \ln 0.5 < 0.5$$

# Exemple : cas Gaussien 2-d



# Exemple : cas Gaussien 2-d

On estime les paramètres des gaussiennes sur les données

- ▶ Estimation de la moyenne :

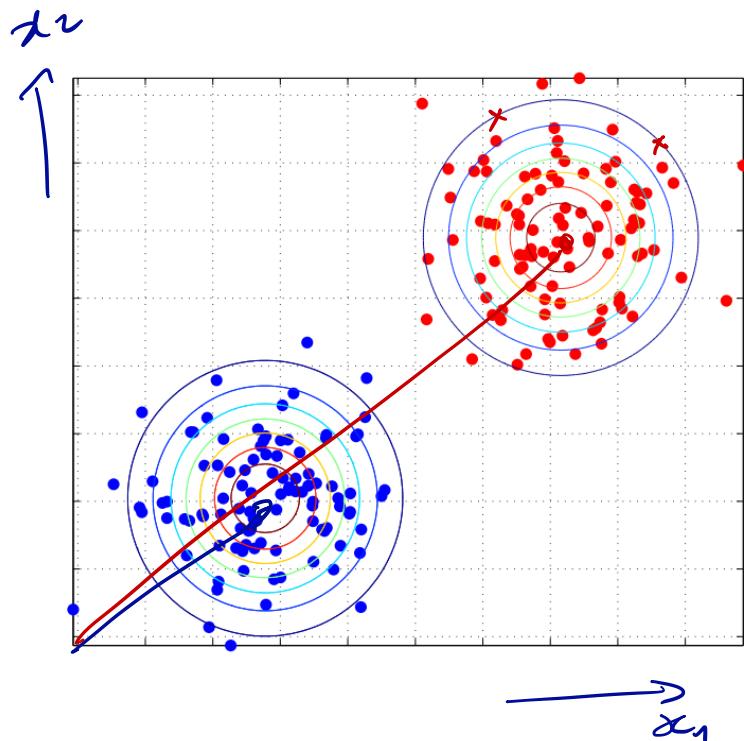
$$\hat{\mu}_i = \frac{1}{N_i} \sum_{m=1}^{N_i} \vec{x}_m$$

- ▶ Estimation de la matrice de covariance :  $\Sigma_i =$

$$\Rightarrow \frac{1}{N_i - 1} \sum_{m=1}^{N_i} (\vec{x}_m - \hat{\mu}_i)^t (\vec{x}_m - \hat{\mu}_i)$$

$$\Sigma_i = \frac{1}{N_i - 1} (\mathbf{X} - \mu_i)^t (\mathbf{X} - \mu_i)$$

$$\mathbf{X} = [\vec{x}_0, \vec{x}_1, \dots, \vec{x}_{n-1}]$$



# Apparté : lignes de niveau

Rappel : densité de probabilité conditionnelle multivariée pour une classe  $c_i$  (ici,  $i$  vaut 1 ou 2) :

$$p(\mathbf{x}|C_i) = \frac{1}{2\pi|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^t \Sigma_i^{-1} (\mathbf{x} - \mathbf{m}_i)\right)$$

## Courbes de niveau ou courbes d'équidensité

Ce sont les courbes qui relient les points  $\mathbf{x}$  pour lesquels  $(\mathbf{x} - \mathbf{m}_i)^t \Sigma_i^{-1} (\mathbf{x} - \mathbf{m}_i) = k$ , où  $k$  est une constante positive

# Exemple : cas Gaussien 2-d

On détermine les frontières de décision :

- ▶ Équation d'une frontière :  $P(x|c_1)P(c_1) = P(x|c_2)P(c_2)$
- ▶ On substitue les  $P(x|c_i)$  par l'expression des Gaussiennes :
  - ▶  $x|c_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$
  - ▶  $x|c_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$

# Exemple : cas Gaussien 2-d

On détermine les frontières de décision :

- ▶ Équation d'une frontière :  $P(x|c_1)P(c_1) = P(x|c_2)P(c_2)$
- ▶ On substitue les  $P(x|c_i)$  par l'expression des Gaussiennes :
  - ▶  $x|c_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$
  - ▶  $x|c_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$

On peut définir une fonction "discriminante"  $g$  telle que :

$x$  est classé dans la classe  $c_i$  si  $g_i(x) > g_j(x)$ , pour tout  $j \neq i$

Les frontières de décision sont alors définies par :

$$g_{ij} \equiv (g_i(x) = g_j(x))$$

# Fonctions discriminantes

- Soit  $g_i$  la fonction correspondant au logarithme népérien de la fonction  $p(\mathbf{x}|C_i) * P(C_i)$ , où  $P(C_i)$  est la probabilité *a priori* de la classe  $C_i$ .
- $g_i$  est utilisée lors de la phase de décision de la classification d'observations entre les deux classes :

$$g_i(\mathbf{x}) = \log(p(\mathbf{x}|C_i)P(C_i))$$
$$= \underbrace{\gamma_i(\mathbf{x})}_{\gamma_i(\mathbf{x}) = -\log(\Sigma_i^{-1}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i))} + \underbrace{\log P(C_i)}_{+\log P(C_i)}$$

**Exercice** Donner l'expression exacte de  $g_i$  lorsque les distributions conditionnelles sont normales, en fonction de  $\mathbf{x}, m_i, \Sigma_i, P(C_i)$  **en distribuant le log**

$$g_i(\mathbf{x}) = \log\left(\frac{1}{2\pi|\Sigma_i|^{1/2}}\right) e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)} \times P(C_i)$$
$$= \log\left(\frac{1}{\sqrt{A}}\right) + \log(P(C_i))$$

# Fonctions discriminantes

- ▶ La frontière entre les 2 classes  $C_1$  et  $C_2$  est donnée par :

$$g_2(\mathbf{x}) - g_1(\mathbf{x}) = 0$$

**Exercice** Simplifier le plus possible cette expression lorsque les déterminants des matrices de covariances sont égaux :

$|\Sigma_1| = |\Sigma_2|$  (**égalité des déterminants des deux matrices de covariances**, mais attention on reste dans le cas général où ces matrices peuvent être différentes).

$$\begin{aligned} & - (\mathbf{x} - \boldsymbol{\mu}_1)^T \underbrace{(\Sigma_1^{-1})}_{=} (\mathbf{x} - \boldsymbol{\mu}_1) + \log P(C_1) \\ &= - (\mathbf{x} - \boldsymbol{\mu}_2)^T \underbrace{(\Sigma_2^{-1})}_{=} (\mathbf{x} - \boldsymbol{\mu}_2) + \log P(C_2) \end{aligned}$$

# Retour à l'exemple : cas Gaussien 2-d

Cas matrices de covariance isotropiques ou sphériques :

$$\Sigma_i = \sigma_i^2 \mathbf{I}_2$$

→ Frontières de décision : droites

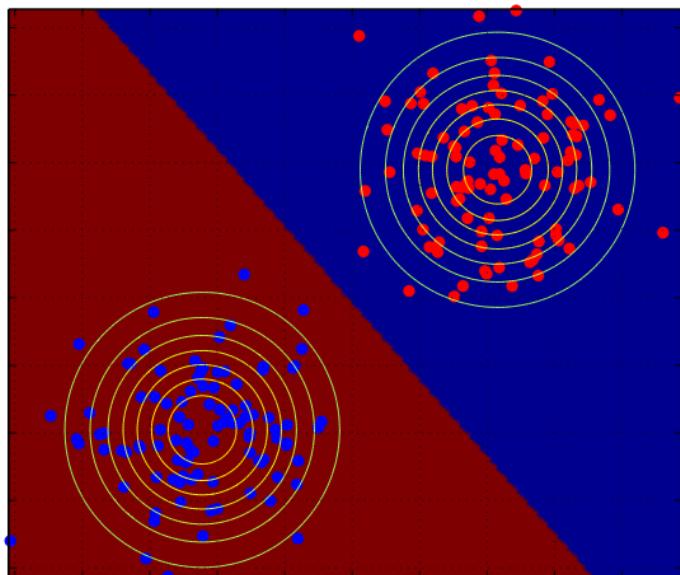
$$\Sigma_1 = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

→ Fonction discriminante

$$g_i(x) = \theta_i^t x + b_i \quad | \text{ droite}$$

$$\theta_i = \mu_i / \sigma_i^2$$

$$b_i = -\frac{\mu_i^t \mu_i}{2\sigma_i^2} + \log P(c_i)$$



# Cas Gaussien 2-d : frontières quadratiques

Cas matrices de covariance quelconques : frontières plus complexes

Fonction discriminante

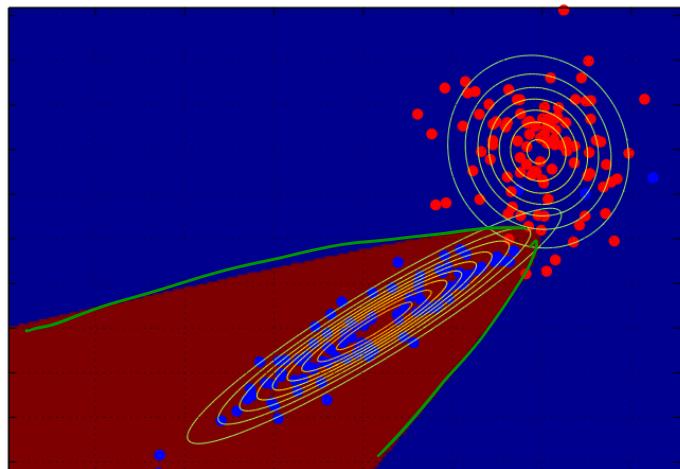
$$g_i(x) = \underline{\mathbf{x}^t \mathbf{W}_i \mathbf{x}} + \underline{\mathbf{w}_i^t \mathbf{x}} + w_i$$

$$\mathbf{W}_i = -\frac{1}{2} \underline{\Sigma_i^{-1}}$$

$$\mathbf{w}_i = \underline{\Sigma_i^{-1} \mu_i}$$

$$w_i = -\frac{1}{2} \underline{\mu_i^t \Sigma_i^{-1} \mu_i} - \underline{\frac{1}{2} \log |\Sigma_i|} + \underline{\log P(c_i)}$$

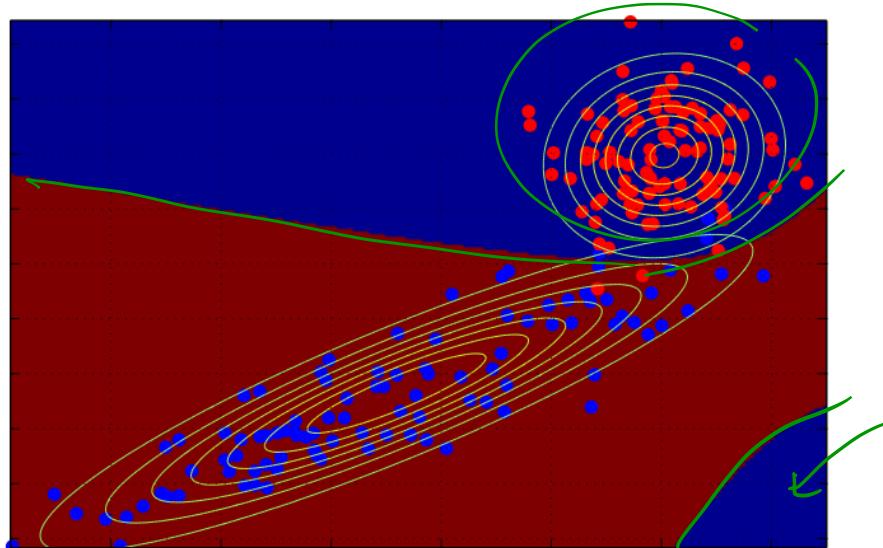
Exemple : frontières paraboliques



# Cas Gaussien 2-d : frontières quadratiques

Cas matrices de covariance quelconques : frontières plus complexes

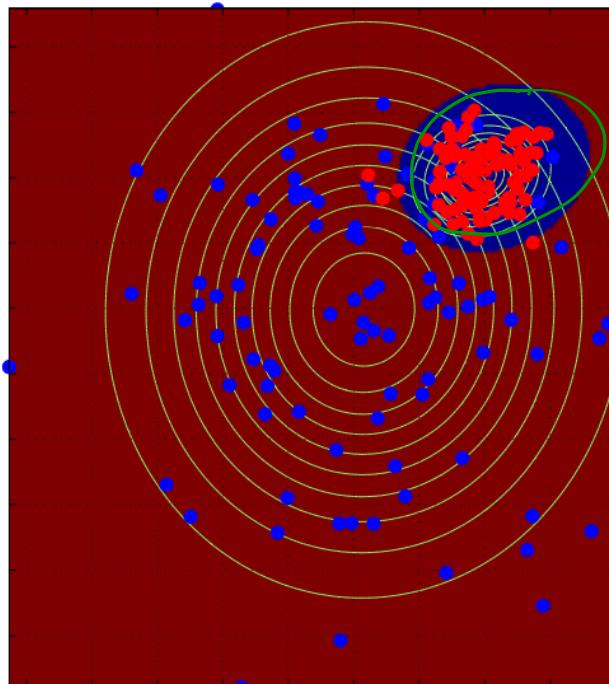
Exemple : frontières hyperboliques



# Cas Gaussien 2-d : frontières quadratiques

Cas matrices de covariance quelconques : frontières plus complexes

Exemple : frontières ellipsoïdes



# Classifieur Bayésien naïf

Pour estimer  $\Sigma_i$ , le nombre de paramètres est  $O(d^2/2)$ , ce qui nécessite un grand nombre d'observation si la dimension  $d$  est grande.

Classifieur Bayésien naïf : on simplifie le problème en supposant que les features sont indépendantes les unes des autres :

paramètres  
x

$$p(\mathbf{x}|c_i) = \prod_{k=1}^d p(x_k|c_i) = p(x_1|c_i) \times p(x_2|c_i)$$