M2 CHPS

[Data & Apprentissage]

Introduction à la science des données et à l'apprentissage

Nicolas Vayatis

Problème de classification, méthodes paramétriques

Données de classification - Exemples

- **Diagnostic médical :**
  - X : résultats des analyses médicales
  - Y : diagnostic,
  - $Y = +1$ si le patient est en bonne santé, et $Y = -1$ sinon

- **Risque de crédit :**
  - X : données socio-économiques d'un individu,
  - Y : indicateur de défaut de paiement,
  - $Y = +1$ si l'emprunteur est fiable, et $Y = -1$ sinon

- **Anti-spam :**
  - X : descripteur d'un email,
  - Y : statut de l'email,
  - $Y = +1$ si le message est un spam, et $Y = -1$ sinon

Quelles décisions pour ces données ?

1. **Classification**

   But : Prédire les nouveaux labels Y

   Satisfaits si l'erreur de classification est faible

2. **Scoring**

   But : Ranger les X dans une liste

   Satisfaits si beaucoup de $Y = +1$ sont en tête de liste

Plan de la séance

1. Modèle probabiliste des *données* de classification
2. Cadre théorique du *problème* de classification
3. Méthodes de classification paramétriques (linéaires) classiques
    1. Analyse discriminante (LDA/QDA)
    2. Analyse discriminante de *Fischer* (FDA)
    3. Régression logistique linéaire
4. De la classification au scoring (ciblage) : la courbe ROC et l'aire AUC
5. Algorithme du perceptron : linéaire et non-paramétrique !

Modèle probabiliste des *données* de classification

Modèle probabiliste pour la classification supervisée

- (X, Y) - couple de variables aléatoires de loi de probabilité inconnue P
- $X \in \mathcal{X}$ - observation sur un espace mesurable (e.g. $\mathbb{R}^d$)
- $Y \in \{-1, +1\}$ - label/classe binaire (par simplicité)

$\longrightarrow$ Description de la loi jointe $P = \mathcal{L}(X, Y)$ du couple aléatoire (X, Y) à partir des lois conditionnelles ?

Description de la loi jointe

1. Approche générative : $\mathcal{L}(X, Y) = \mathcal{L}(Y) \otimes \mathcal{L}(X \mid Y)$
   - Modèle de type mélange de paramètre : $p = \mathbb{P}\{Y = +1\} \in [0, 1]$
   - Lois conditionnelles sur $\mathbb{R}^d$:

$$P_+ = \mathcal{L}(X \mid Y = +1) \qquad \text{et} \qquad P_- = \mathcal{L}(X \mid Y = -1)$$

2. Approche discriminative $\mathcal{L}(X, Y) = \mathcal{L}(X) \otimes \mathcal{L}(Y \mid X)$
   - Loi marginale sur $\mathbb{R}^d$: $P_X = \mathcal{L}(X)$
   - Meilleure prévision :

$$\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}\ , \quad \forall x \in \mathbb{R}^d$$

Lien entre les deux descriptions

- Loi marginale (les "dP" désignent les densités des lois) :

$$dP_X = p dP_+ + (1-p) dP_-$$

- Probabilité a posteriori (fonction de régression) :

$$\forall x \in \mathcal{X} \ , \qquad \eta(x) = \frac{p \ dP_+}{p \ dP_+ + (1-p) \ dP_-}(x)$$

- Une remarque :

$$\eta(x) > \frac{1}{2} \quad \Leftrightarrow \quad p \ dP_+(x) > (1-p) \ dP_-(x)$$

Cadre théorique du *problème* de classification

Le problème de classification binaire

- **Données disponibles :** $(x_1, y_1), \ldots, (x_n, y_n)$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
- **Problème :** prédiction du label y connaissant x
- **On cherche :** un classifieur $g : \mathbb{R}^d \to \{-1, +1\}$
- **Question :** trouver un classifieur g qui "généralise" bien.
- **Idée :** on choisit g qui "interprète" bien, mais pas trop !
- **Concrètement :** souvent on cherche une fonction de décision $f : \mathbb{R}^d \to \mathbb{R}$ et on y associe le classifieur $g = \mathrm{sgn}(f)$

Critère d'évaluation : erreur d'un classifieur

- Etant donnée une observation x, un classifieur g réalise une prédiction $g(x)$ à comparer à la classe y.

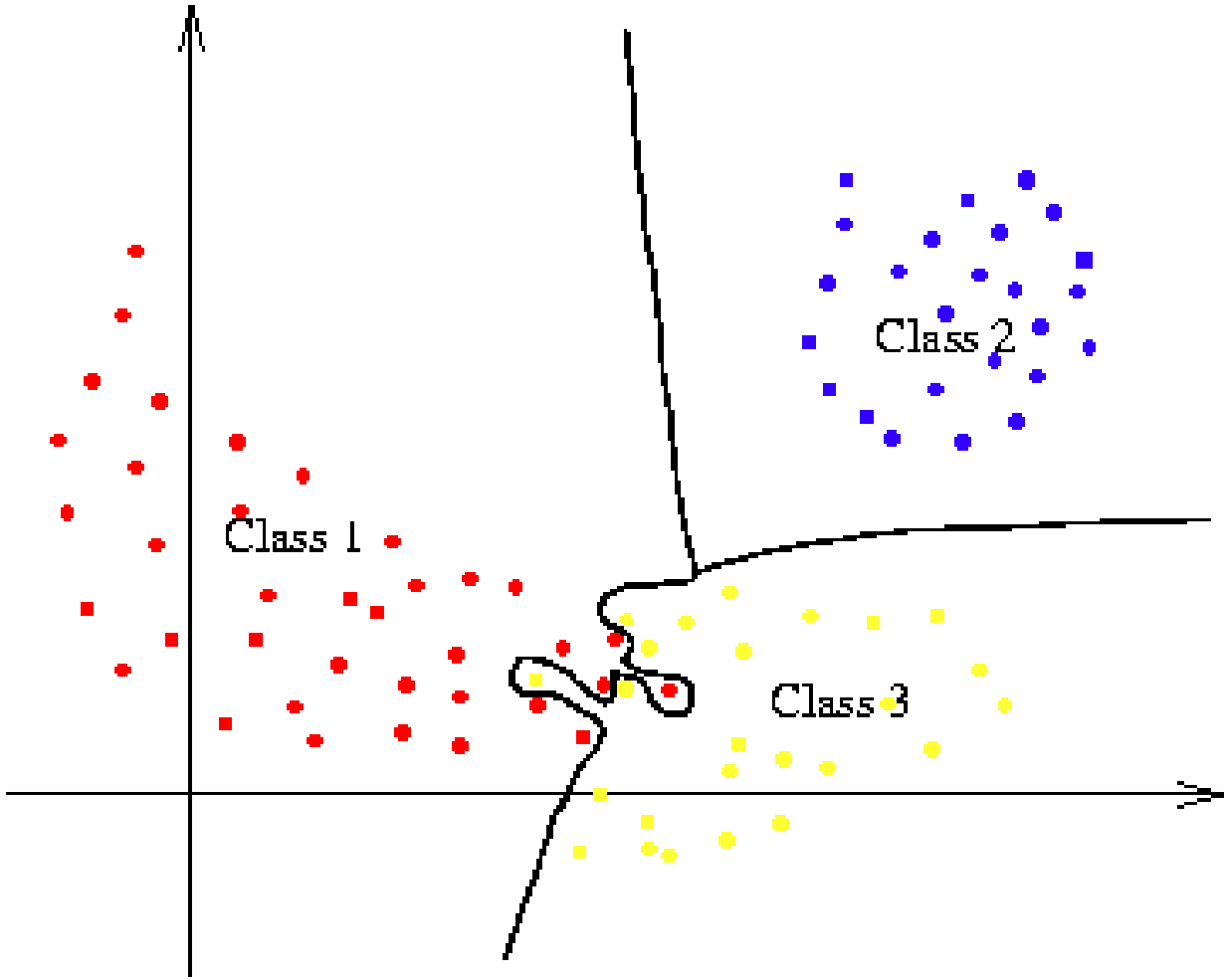
Erreur du classifieur = Taux d'observations mal classées

$$\hat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[g(x_i)\neq y_i]} = \frac{\#\{i \ : \ g(x_i) \neq y_i\}}{n}$$

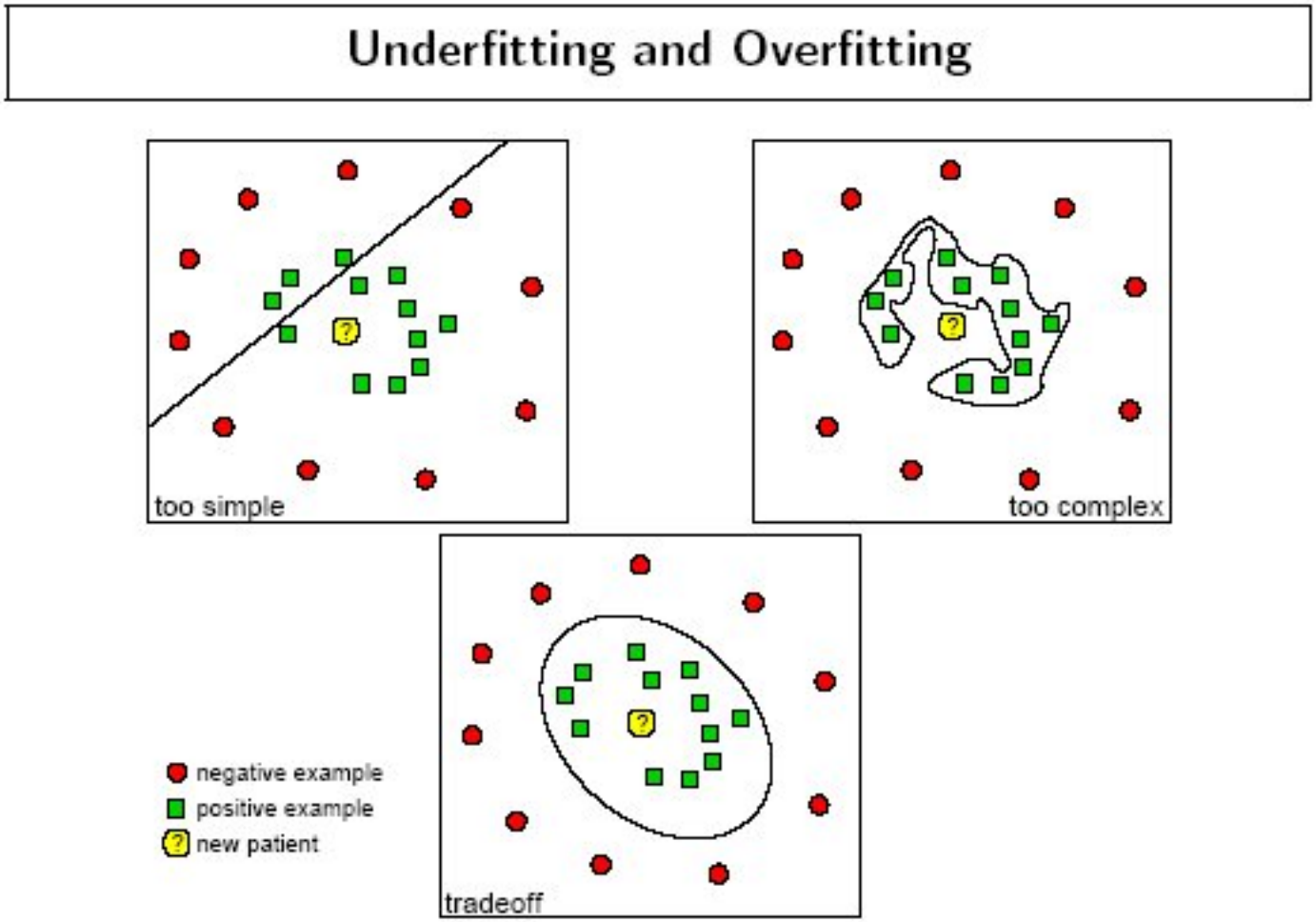
Cette erreur s'appelle aussi erreur d'apprentissage.

- Un classifieur g "interprète" convenablement les données si son erreur d'apprentissage est faible.
- Attention ! si on s'intéresse seulement à ce type d'erreur, on risque d'avoir des problèmes...

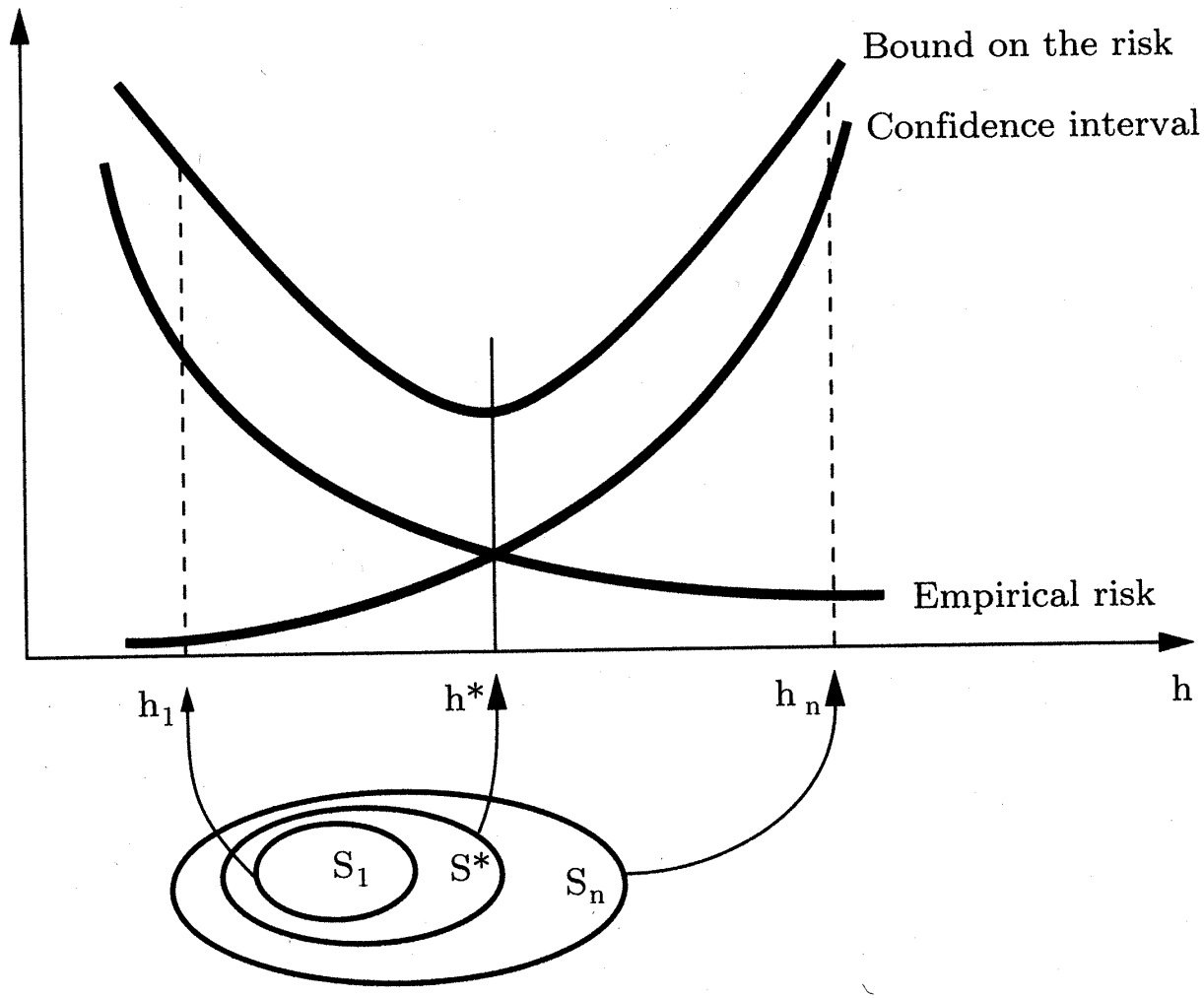
Overfitting en classification



Overfitting en classification (suite)



Calibration de la complexité



En pratique : Stratégie du holdout

- On sépare les données disponibles en deux sous-groupes :
  - Base d'apprentissage : $(X_1, Y_1), \ldots, (X_n, Y_n)$
  - Base de test : $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$
- L'erreur de test $\hat{L}'_m(g) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{[g(X_{n+j}) \neq Y_{n+j}]}$ est une estimation de $L(g)$ pour tout classifieur g (on peut conditionner par rapport à la base d'apprentissage si g résulte d'un apprentissage).
- Meilleure pratique : la *validation croisée* pour rendre plus robuste l'estimateur de $L(g)$

Méthode # 1 : Analyse discriminante linéaire (LDA) et quadratique (QDA)

Rappel : loi gaussienne multivariée

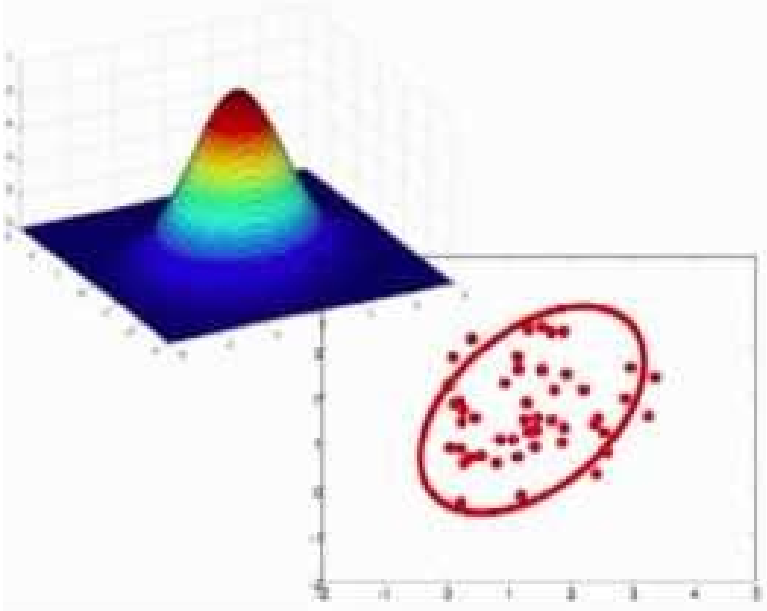
Multivariate Gaussian models

- Similar to univariate case

$$\mathcal{N}(\underline{x}\,;\,\underline{\mu},\Sigma) = \frac{1}{(2\pi)^{d/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})\Sigma^{-1}(\underline{x}-\underline{\mu})^T\right\}$$

$\underline{\mu}$ = length-d row vector
Σ = d x d matrix

$|\Sigma|$ = matrix determinant



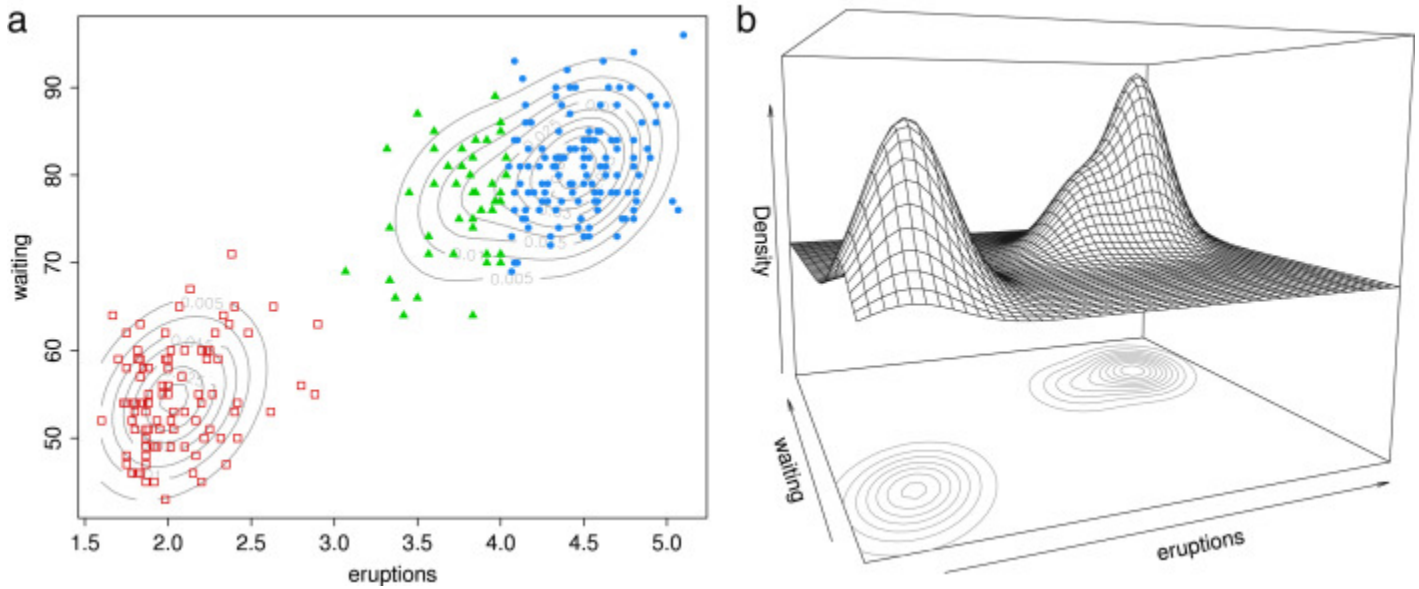
Maximum likelihood estimate:

$$\hat{\underline{\mu}} = \frac{1}{m}\sum_j \underline{x}^{(j)}$$

$$\hat{\Sigma} = \frac{1}{m}\sum_j (\underline{x}^{(j)} - \hat{\underline{\mu}})^T(\underline{x}^{(j)} - \hat{\underline{\mu}})$$

(average of dxd matrices)

Rappel : mélange gaussien en 2D



Hypothèse : le modèle paramétrique de mélange gaussien

- $X \in \mathbb{R}^d$ et $Y \in \{1, \ldots, K\}$
- Forme paramétrique gaussienne pour la loi a posteriori

$$\mathbb{P}(X \mid Y = k) \sim \mathcal{N}(m_k, \Sigma_k) \ , \qquad \text{densité } f_k$$

- Paramètre de mélange π_k pour la classe $Y = k$
- On exprime alors :

$$\eta_k(x) = \mathbb{P}(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$$

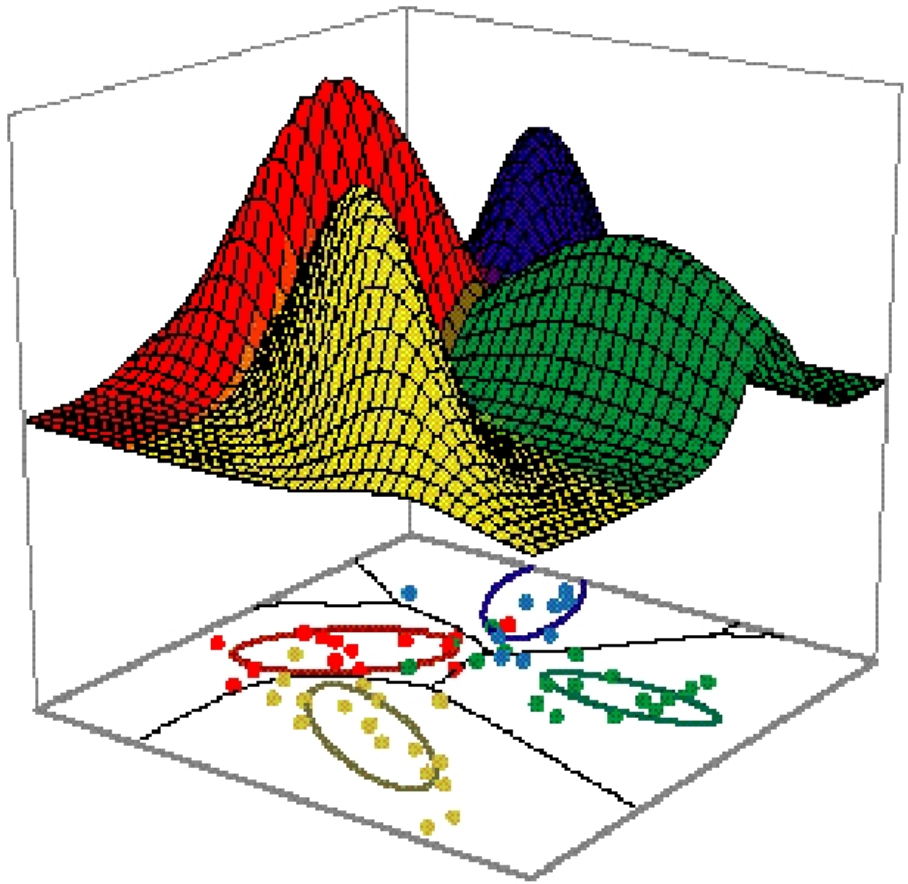
Analyse discriminante linéaire (LDA)

- On suppose que $\Sigma_k = \Sigma$, $\forall k$
- On exprime alors :

$$\begin{aligned}
\log\left(\frac{\eta_k(x)}{\eta_j(x)}\right) &= \frac{\pi_k f_k(x)}{\pi_j f_j(x)} \\
&= \log\left(\frac{f_k(x)}{f_j(x)}\right) + \log\left(\frac{\pi_k}{\pi_j}\right) \\
&= -\frac{1}{2}(m_k + m_j)^T \Sigma^{-1}(m_k + m_j) \\
&\quad + \log\left(\frac{\pi_k}{\pi_j}\right) + x^T \Sigma^{-1}(m_k - m_j)
\end{aligned}$$

- Equation linéaire en x !

Analyse discriminante linéaire (suite)



Analyse discriminante quadratique (QDA)

- Cas où les matrices Σ_k, $\forall k$ sont distinctes
- On obtient dans ce cas des fonctions discriminantes :

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x - m_k)^T \Sigma_k^{-1}(x - m_k) \\ &\quad + \log(\pi_k) - \frac{1}{2}\log det(\Sigma_k)\end{aligned}$$

- Séparatrices quadratiques en x !

Lien entre LDA et QDA

- Estimation des matrices Σ_k coûteuse en grande dimension
- Alors : LDA avec termes de couplage ou QDA ?
- Idée : régulariser la matrice par interpolation

$$\hat{\Sigma}_k(\lambda) = \lambda \hat{\Sigma}_k(\lambda) + (1 - \lambda)\hat{\Sigma}$$

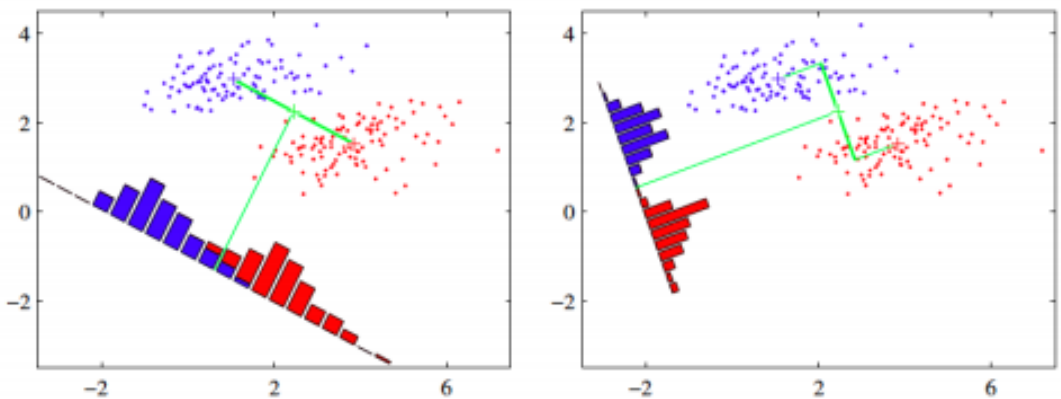
  cf. Friedman (1989)

- Variations autour de régularisation et parcimonie

Méthode # 2 : Analyse discriminante de Fisher (FDA)

Principe de l'analyse discriminante de Fisher

- Hypothèse : On considère deux lois gaussiennes pour $\mathcal{L}(X \mid Y)$
- Heuristique : On considère le séparateur linéaire qui maximise la distance entre les centres des lois projetées normalisée par la variance totale des projections sur le vecteur normal au séparateur.



Plus formellement : Notations pour $i = 1, 2$

- Gaussiennes $\mathcal{N}(\mu_i, \Sigma_i)$ échantillonnées en données de classification binaires
- Estimateurs empiriques des paramètres $\hat{\mu}_i$, $\hat{\Sigma}_i$
- Centres projetés sur un vecteur $u \in \mathbb{R}^d$: $m_i(u) = u^T \hat{\mu}_i$
- Dispersion des observations projetées :

$$\hat{S}_i^2(u) = \sum_{j : Y_j = i} (u^T X_j - m_i(u))^2$$

Formulation de la FDA

- Critère à maximiser pour $u \in \mathbb{R}^d$:

$$J(u) = \frac{(m_1(u) - m_2(u))^2}{\hat{S}_1^2(u) + \hat{S}_2^2(u)} = \frac{u^T S_B u}{u^T S_W u}$$

où S_B et S_W peuvent être interprétées comme des matrices de dispersion inter et intra-classes.

- La solution par lagrangien est obtenue via la résolution du problème aux valeurs propres :

$$S_B u = \lambda S_W u$$

Si S_W de rang plein, alors la solution est explicite :
$u = S_W^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$

- A noter que la direction u n'est pas liée à la direction obtenue pour la PCA.

Méthode # 3 : Régression logistique linéaire

Modèle statistique

- On a : $Y \in \{1, \ldots, K\}$, $X \in \mathbb{R}^d$
- On pose $\eta_k(x) = \mathbb{P}\{Y = k \mid X = x\}$ pour $k \in \{1, \ldots, K\}$
- On suppose que $\forall k$, qu'il existe $\theta_k \in \mathbb{R}^d$ tel que

$$\log\left(\frac{\eta_k(x)}{\eta_K(x)}\right) = \theta_k^T x$$

- Ou bien :

$$\eta_k(x) = \frac{\exp(\theta_k^T x)}{1 + \sum_{j=1}^{K-1} \exp(\theta_j^T x)}$$

et $\theta_K = 1$.

Fit d'un modèle de régression logistique

- On note $\theta = (\theta_1, \ldots, \theta_{K-1})$ et $\eta_k(x) = p_k(x, \theta)$
- Log-vraisemblance

$$\ell(\theta) = \sum_{i=1}^{n} \log p_k(X_i, \theta)$$

- Cas où $K = 2$, $\theta \in \mathbb{R}^d$, $p(x, \theta) = p_1(x, \theta)$

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \big(Y_i \log p(X_i, \theta) + (1 - Y_i) \log(1 - p(X_i, \theta)) \big) \\ &= \sum_{i=1}^{n} \big(Y_i \theta^T X_i - \log(1 + \exp(\theta^T X_i)) \big) \end{aligned}$$

Résolution numérique

- Equation de score

$$\frac{\partial \ell}{\partial \theta}(\theta) = \sum_{i=1}^{n} X_i \big(Y_i - p(X_i, \theta) \big) = 0$$

- Matrice hessienne

$$H_\ell(\theta) = -\sum_{i=1}^{n} X_i X_i^T \, p(X_i, \theta) \big(1 - p(X_i, \theta) \big)$$

- Schéma itératif de Newton-Raphson

$$\theta_{t+1} = \theta_t - \left(H_\ell(\theta_t) \right)^{-1} \frac{\partial \ell}{\partial \theta}(\theta_t)$$

- Se ramène à un estimateur des moindres carrés pondéré...

Limites des méthodes génératives

- Modèles statistiques paramétriques : "All models are wrong"...
- Lourd a priori de modélisation : "... some are useful"
  - cadre gaussien
  - modèle linéaire
- Curse of dimensionality (cf. Bellmann)

De la classification au scoring : Courbe ROC et AUC

Deux types d'erreur

- Décomposition de l'erreur de classification

$$L(g) = \mathbb{P}\{g(X) = +1, Y = -1\} + \mathbb{P}\{g(X) = -1, Y = +1\}$$

- Taux de faux positifs

$$\alpha(g) = \mathbb{P}\{g(X) = +1 \mid Y = -1\}$$

- Taux de vrais positifs

$$\beta(g) = \mathbb{P}\{g(X) = +1 \mid Y = +1\}$$

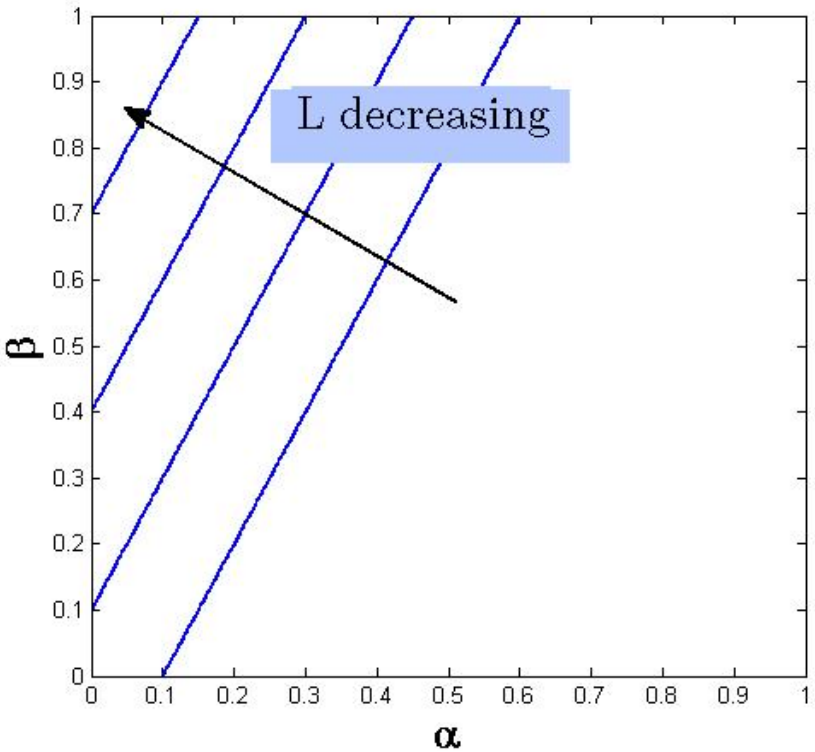
- On remarque que :

$$L(g) = \mathbb{P}\{Y \neq g(X)\} = (1-p)\alpha(g) + p(1-\beta(g))$$

Diagramme α-β

- Pour une proportion p fixée, une erreur de classification fixée $L(g) = L$, on a :

$$\beta = \left(\frac{1-p}{p}\right)\alpha + 1 - \frac{L}{p}$$



Tests d'hypothèses

- Sous l'observation X, tester

$$H_0 : Y = -1 \quad \text{contre} \quad H_1 : Y = +1$$

- Statistique de test optimale (Neyman-Pearson)

$$T^*(X) = \frac{1-p}{p} \cdot \frac{\eta(X)}{1-\eta(X)}$$

- $\alpha =$ Erreur de première espèce
- $\beta =$ Puissance du test

Classifieur de Neyman-Pearson

- Pour α fixé, la région de rejet est :

$$R_\alpha^* = \{\ x\ :\ \eta(x) > Q^-(\eta, \alpha)\ \}$$

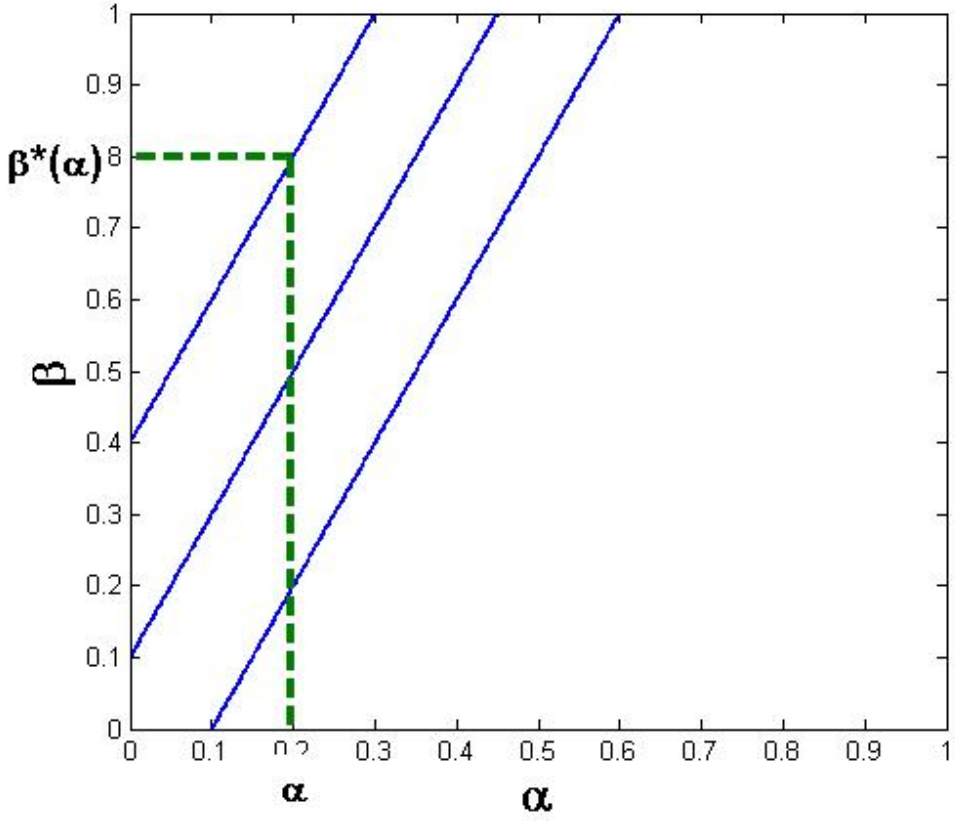
où $Q^-(\eta, \alpha) = (1-\alpha)$-quantile de $\mathcal{L}(\eta(X) \mid Y = -1)$

- Soit le classifieur :

$$g_\alpha^*(x) = 2\mathbb{I}\{\ x \in R_\alpha^*\ \} - 1$$

Classifieur NP optimal

- En général : $L(g_\alpha^*) > L^*$ sauf si $Q^-(\eta, \alpha) = 1/2$
- Soit $\beta^*(\alpha) = \beta(g_\alpha^*)$



Notations
Performance d'une règle de scoring

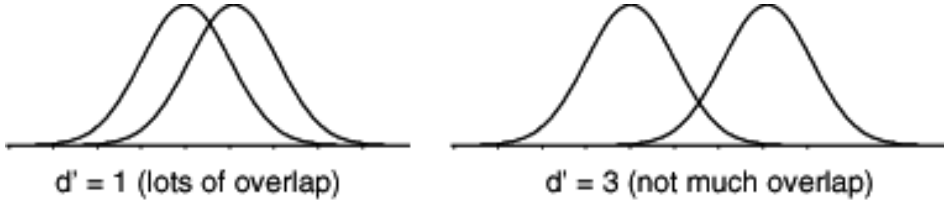
- Consider $s : \mathbb{R}^d \to \mathbb{R}$ a detector response (socring rule)
- A hit corresponds to $Y = +1$, an alarm to $\{s(X) \geq t\}$
- True positive rate and false positive rate :

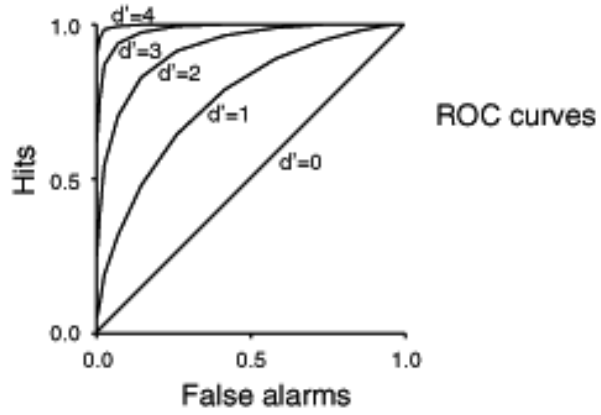
$$\begin{aligned} \beta(s,t) = &\quad \mathbb{P}\{s(X) \geq t \mid Y = +1\} \quad (\text{TPR}) \to \max \\ \alpha(s,t) = &\quad \mathbb{P}\{s(X) \geq t \mid Y = -1\} \quad (\text{FPR}) \to \min \end{aligned}$$

- Main point : trade-off required since

$$\begin{aligned} \beta(s,t) \to 1 \quad &\text{but} \quad \alpha(s,t) \to 1 \quad \text{when} t \to -\infty \\ \alpha(s,t) \to 0 \quad &\text{but} \quad \beta(s,t) \to 0 \quad \text{when} t \to +\infty \end{aligned}$$

Courbes ROC idéales





- Pour $s : \mathbb{R}^d \to \mathbb{R}$ règle fixée
- Courbe ROC d'une règle de scoring s :

$$t \in \mathbb{R} \mapsto (\alpha_s(t), \beta_s(t))$$

Optimal elements for scoring

- $X \in \mathbb{R}^d$ - observation vector in a high dimensional space
- $Y \in \{-1, +1\}$ - binary diagnosis (i.e. classification data)
- Key theoretical quantity (posterior probability)

$$\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\} \ , \quad \forall x \in \mathbb{R}^d$$

- Optimal scoring rules :

  $\Rightarrow$ increasing transformations of η

Critère pratique - Aire sous la courbe ROC (AUC)

- Pour toute règle de scoring s, soit :

$$\begin{aligned}\mathrm{AUC}(s) &= \int_0^1 \mathrm{ROC}(s,\alpha)\, d\alpha \\ &= \mathbb{P}\{s(X) > s(X') \mid Y > Y'\} \\ &\quad + \frac{1}{2}\mathbb{P}\{s(X) = s(X') \mid Y > Y'\}\end{aligned}$$

où $(X, Y), (X', Y')$ i.i.d.

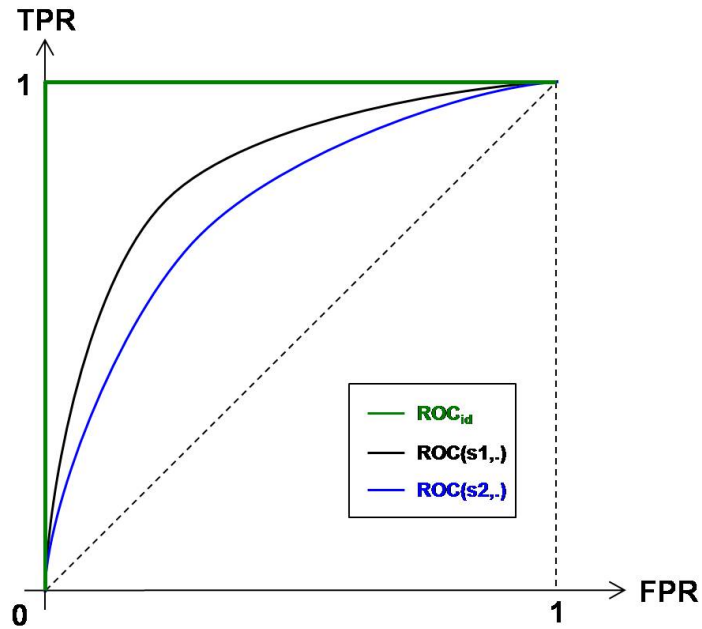
- AUC maximale

$$\mathrm{AUC}^* = \mathrm{AUC}(\eta) = \frac{1}{2} + \frac{\mathbb{E}(|\eta(X) - \eta(X')|)}{4p(1-p)},$$

- La convergence au sens de l'AUC correspond à la convergence L_1 des courbes ROC

Performance measures for scoring

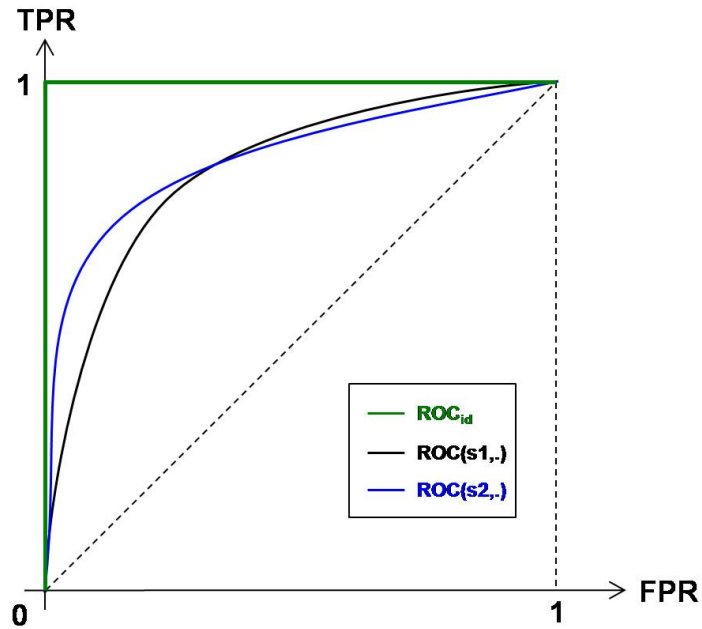
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



ROC curves.

Performance measures for scoring

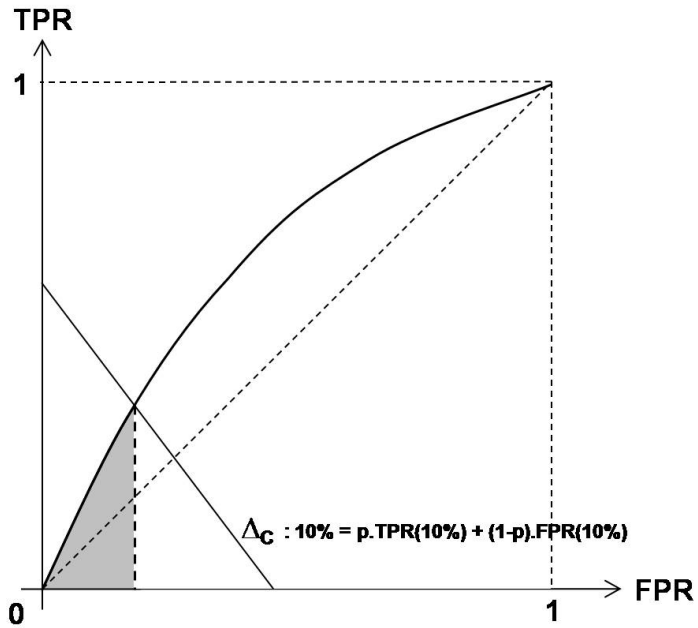
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



ROC curves.

Performance measures for scoring

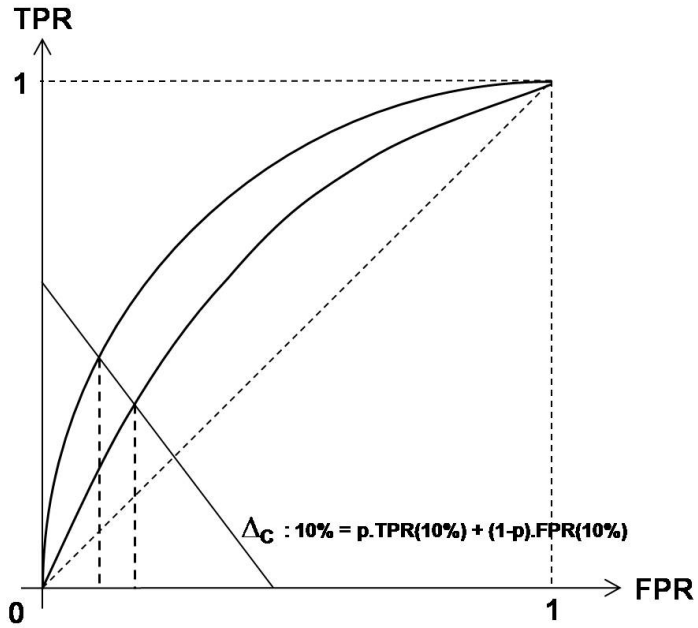
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



Partial AUC.

Performance measures for scoring

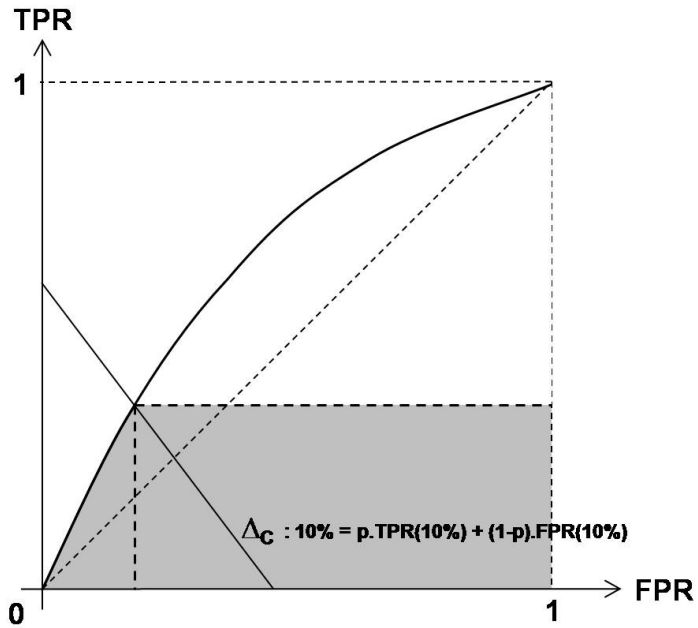
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



Inconsistency of Partial AUC.

Performance measures for scoring

- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



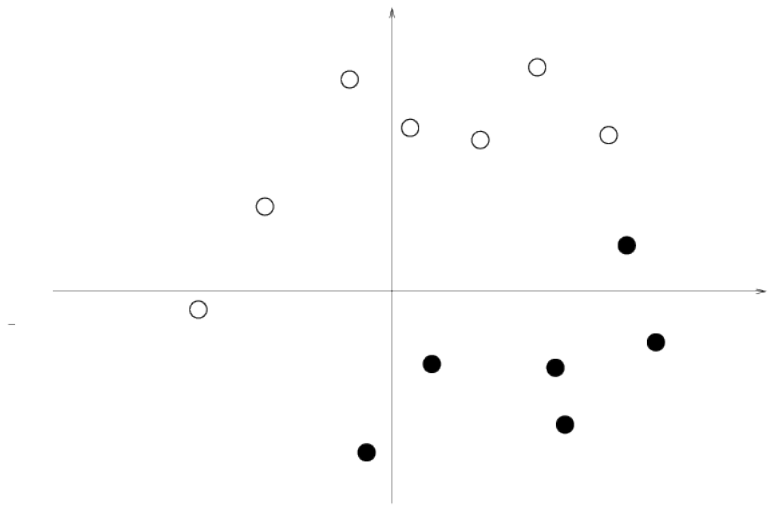
Local AUC.

Algorithmes de discrimination linéaire *non-paramétrique*

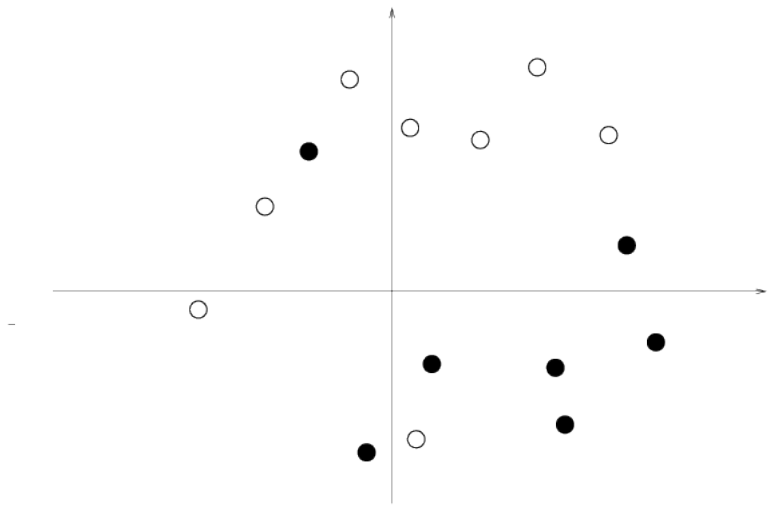
Les trois scénarios

1. les populations sont linéairement séparables
2. les populations sont *presque* linéairement séparables
3. les populations ne sont pas linéairement séparables

Séparabilité linéaire



Scénario 1



Scénario 2

Scénario 1 - Séparateurs linéaires

- **Forme des fonctions de décision :**

$$f(x) = b + < \beta, x >$$

où $b \in \mathbb{R}$, $\beta \in \mathbb{R}^d$.

- L'équation $f(x) = 0$ définit un **hyperplan séparateur** H dans $\mathbb{R}^d$

- **Classifieur associé :**

$$\forall x \in \mathbb{R}^d \quad g_f(x) = \begin{cases} +1 & \text{si } f(x) > 0 \\ \\ -1 & \text{si } f(x) \leq 0 \end{cases}$$

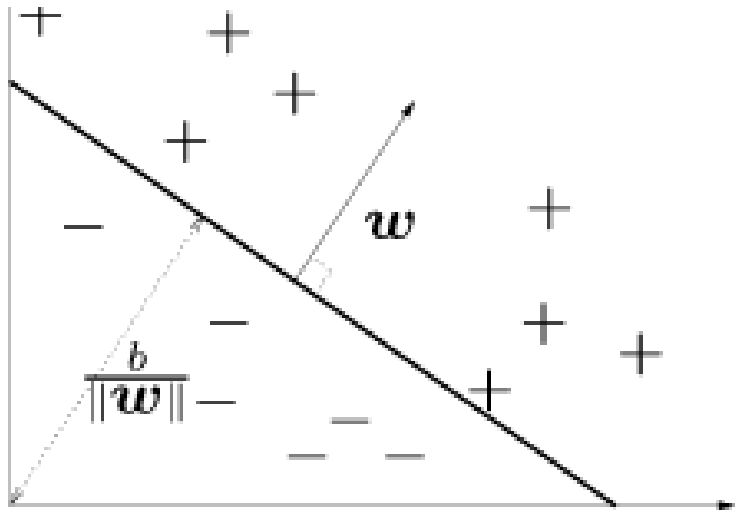
Quelques propriétés

1. $\beta^* = \dfrac{\beta}{\|\beta\|}$ est le vecteur normal à H
2. $\forall x_0 \in H, \quad < \beta, x_0 >= -b$
3. la distance signée (éventuellement négative !) d'un point $x \in \mathbb{R}^d$ à H est donnée par

$$d(x, H) =< \beta^*, x - x_0 >= \frac{1}{\|\beta\|}\left(b + < \beta, x >\right)$$

où $x_0 \in H$

Scénario 1 - Une figure



A separating hyperplane $(\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ for a 2D training set.

Attention ! ici $w = \beta$...

Algorithme du perceptron (Rosenblatt, 1958)

Perceptron - version simplifiée $b = 0$

Génère une suite $\beta_0, \ldots, \beta_n$ de valeurs pour β

1. **Initialisation** - $\beta_0 = 0$
2. **Etape i** - on considère le couple (x_i, y_i) et on regarde s'il est correctement classé ou non

$$\beta_i = \begin{cases} \beta_{i-1} & \text{si } y_i \cdot < \beta_{i-1}, x_i > > 0 \\ \beta_{i-1} + y_i x_i & \text{si } y_i \cdot < \beta_{i-1}, x_i > \leq 0 \end{cases}$$

Algorithme du perceptron général

Perceptron - version générale

- **Paramètres :**
  - taux d'apprentissage η
  - rayon des observations $R = \max_{1 \leq i \leq n} \|x_i\|$

- **Algorithme :**
  1. **Initialisation** - $\beta_0 = 0,\ b_0 = 0$
  2. **Etape i** - si (x_i, y_i) est mal classé par l'hyperplan (b_{i-1}, β_{i-1}), alors :

$$\begin{aligned} \beta_i &= \beta_{i-1} + \eta y_i x_i \\ b_i &= b_{i-1} + \eta y_i^2 R^2 \end{aligned}$$

  sinon $\beta_i = \beta_{i-1},\ b_i = b_{i-1}$

Propriétés du perceptron

Théorème de Novikoff

Si les populations sont linéairement séparables alors l'algorithme du perceptron converge en un nombre fini $T \leq n$ d'étapes où :

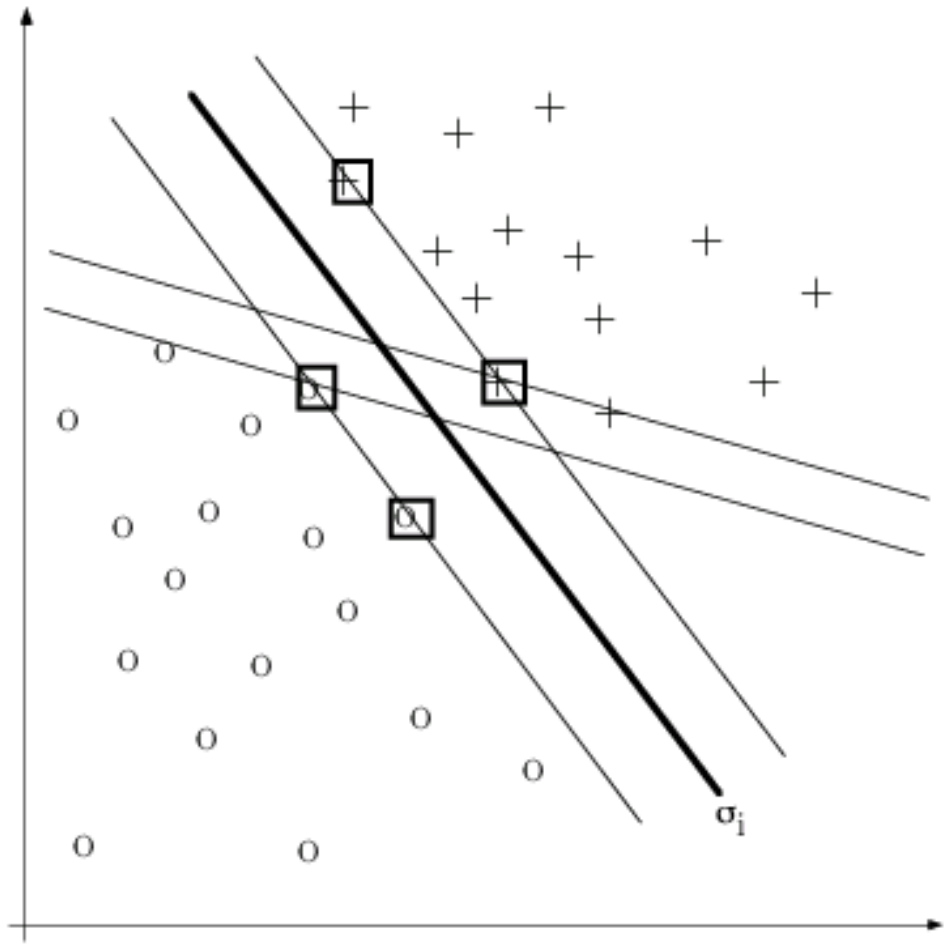
$$T \leq \frac{2R^2}{M^2}$$

avec $M = \min_{1 \leq i \leq n} \{y_i d(x_i, H^*)\}$ pour un certain séparateur H^*.

- **Défaut du perceptron :** mauvaise généralisation
- **Vertu du perceptron :** algorithme séquentiel (online)

Scénario 1 - hyperplan à bonne généralisation

Question : hyperplan se trouvant à distance maximale de chaque population ?



Hyperplan à marges optimales (Vapnik-Chervonenkis, 1964)

Problème d'optimisation

$$\max_{\beta \in \mathbb{R}^d,\ b \in \mathbb{R}} M$$

sous les contraintes :

$$\forall i = 1, \ldots, n\ , \quad y_i \cdot d(x_i, H) \geq M$$

On rappelle :

$$d(x_i, H) = \frac{1}{\|\beta\|}\left(b + < \beta, x_i >\right)$$

Problème quadratique

Contraintes :

$$\forall i = 1, \ldots, n \ , \quad y_i \cdot \frac{1}{\|\beta\|}\big(b + < \beta, x_i >\big) \geq M$$

On peut très bien poser : $M = 1/\|\beta\|$

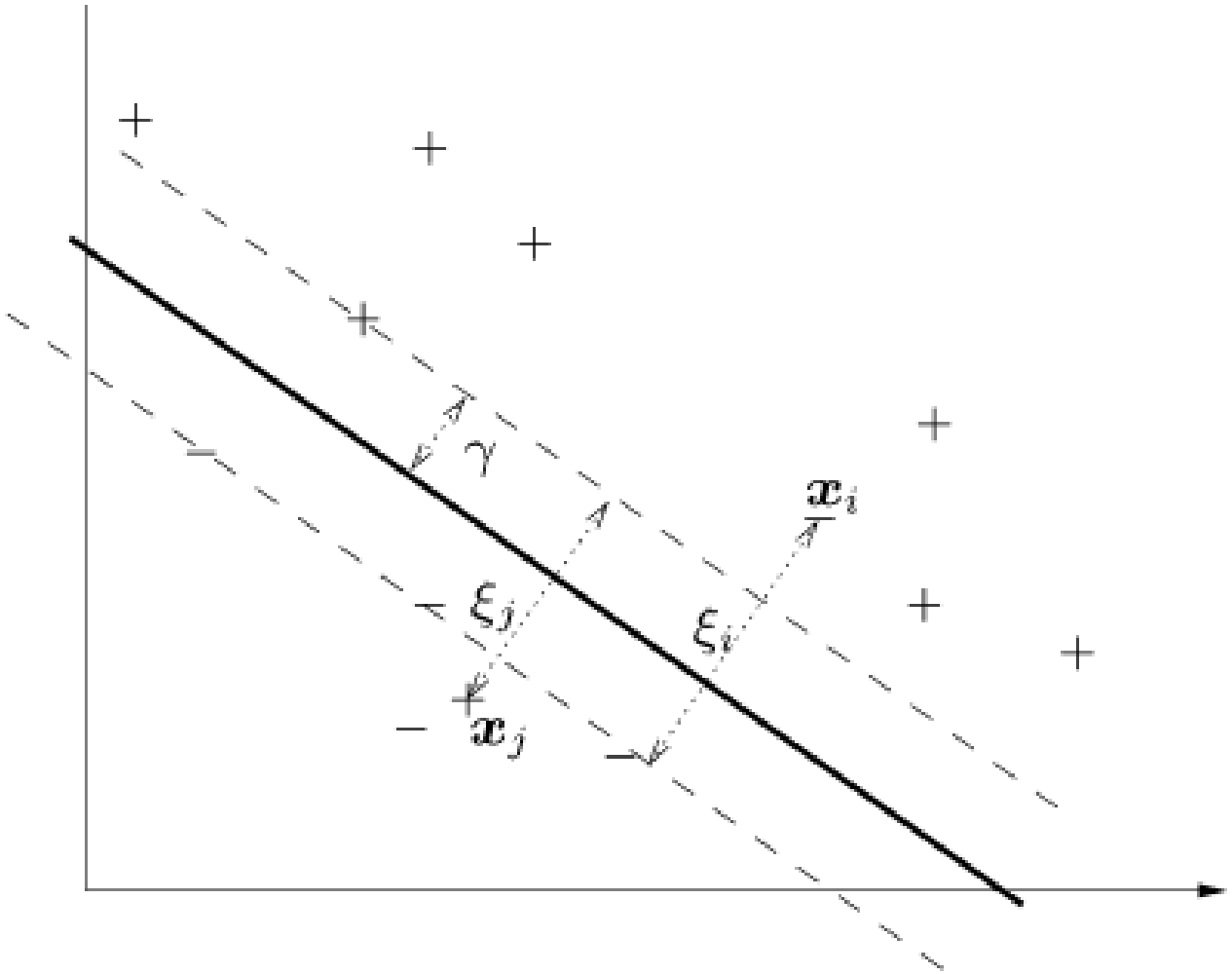
Formulation équivalente

$$\min_{\beta, b} \frac{1}{2}\|\beta\|^2$$

sous les contraintes :

$$\forall i = 1, \ldots, n \ , \quad y_i \cdot \big(b + < \beta, x_i >\big) \geq 1$$

Scénario 2 - Variables "ressorts"



Scénario 2 - Variables "ressorts" (suite)

On introduit n variables supplémentaires ("slacks" ou "ressorts") : $\xi = (\xi_1, \ldots, \xi_n)$ avec $\xi_i \geq 0, \forall i$

Nouveau problème d'optimisation

$$\min_{\beta, b, \xi} \frac{1}{2}\|\beta\|^2$$

sous les contraintes :

$$\begin{aligned} \forall i = 1, \ldots, n\ , \quad y_i \cdot \left(b + < \beta, x_i >\right) &\geq 1 - \xi_i \\ \xi_i &\geq 0 \\ \sum_{i=1}^{n} \xi_i &\leq \Xi \end{aligned}$$

Formulation lagrangienne I

$$\min_{\beta,b,\xi} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} \xi_i$$

sous les contraintes :

$$\begin{aligned} \forall i = 1,\ldots,n\ , \quad \xi_i &\geq 0 \\ \xi_i &\geq 1 - [y_i \cdot \left(b + < \beta, x_i >\right)] \end{aligned}$$

Formulation lagrangienne II

Multiplicateurs de Lagrange : $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\mu = (\mu_1, \ldots, \mu_n)$

Formulation lagrangienne II

$$\min_{\beta,b,\xi} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\big(y_i\cdot\big(b+<\beta,x_i>\big)-(1-\xi_i)\big)+\sum_{i=1}^{n}\mu_i\xi_i$$

Conditions du premier ordre (gradient nul)

$$\begin{aligned} \beta &= \sum_{i=1}^{n}\alpha_i y_i x_i \\ \sum_{i=1}^{n}\alpha_i y_i &= 0 \\ \forall i = 1,\ldots,n\ , \qquad \alpha_i &= C + \mu_i \end{aligned}$$

Formulation duale

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j < x_i, x_j >$$

sous les contraintes :

$$\begin{aligned} \forall i = 1, \ldots, n \ , \qquad 0 &\leq \alpha_i \leq C \\ \sum_{i=1}^{n} \alpha_i y_i &= 0 \end{aligned}$$

On note $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)$ la solution de ce problème.

Conditions de Karush-Kuhn-Tucker

$$\forall i = 1, \ldots, n \ , \qquad \begin{aligned} \alpha_i \big(y_i \cdot f(x_i) - (1 - \xi_i)\big) &= 0 \\ y_i \cdot f(x_i) - (1 - \xi_i) &\geq 0 \\ \alpha_i + \mu_i &= C \\ \mu_i \xi_i &= 0 \\ \beta &= \sum_{i=1}^{n} \alpha_i y_i x_i \\ \sum_{i=1}^{n} \alpha_i y_i &= 0 \end{aligned}$$

Interprétation

Lien entre les coefficients et la position des observations

- si $\hat{\alpha}_i = 0$ alors $y_i \cdot f(x_i) \geq 1 \Rightarrow$ le point x_i est bien classé
  car $\mu_i = C > 0$ et on a $\xi_i = 0$
- si $0 < \hat{\alpha}_i < C$ alors $y_i \cdot f(x_i) = 1 \Rightarrow$ le point x_i est sur la frontière de la marge
  car $\mu_i > 0$ et $\xi_i = 0$
- si $\hat{\alpha}_i = C$ alors $y_i \cdot f(x_i) \leq 1 \Rightarrow$ le point x_i dépasse la frontière de la marge
  car $\mu_i = 0$ et donc $\xi_i \geq 0$

Phénomène remarquable !

En pratique, beaucoup de $\hat{\alpha}_i$ sont nuls !

Solution du problème

Définition

Les $\hat{\alpha}_i \neq 0$ correspondent aux vecteurs de support. On note I l'ensemble des indices parmi $\{1, \ldots, n\}$ correspondants.

Représentation de la solution

Fonction de décision :

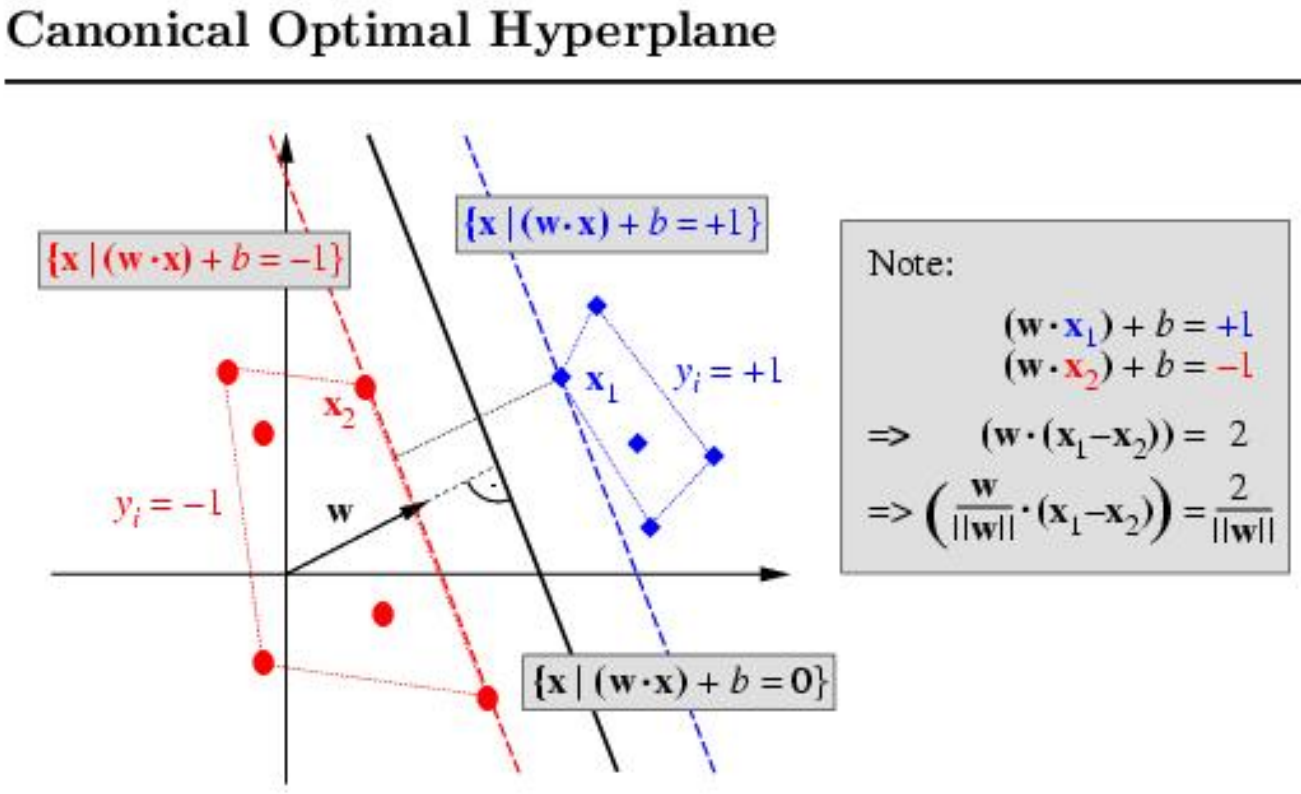
$$\hat{f}(x) = \hat{b} + \sum_{i \in I} \hat{\alpha}_i y_i < x_i, x >$$

où :

$$\hat{\beta} = \sum_{i \in I} \hat{\alpha}_i y_i x_i \ , \quad I = \{i \ : \ \hat{\alpha}_i \neq 0\}$$

$$\hat{b} = y_j - \sum_{i \in I} \hat{\alpha}_i y_i < x_i, x_j > \ , \quad \text{pour un certain } j \in I$$

Vecteurs de support



$\Rightarrow$ Représentation parcimonieuse ("sparse") des SVM

Prochains épisodes

- Classification non supervisée (sans labels)
- Autres algorithmes de classification, non-paramétriques, non-linéaires
- Calibration de la complexité et régularisation de problèmes d'optimisation