# Techniques d’apprentissage supervisé pour discrimination

Séparateurs à Vaste Marge (SVM) – Présentation et applications

# Plan

- SVM
  - Historique
  - Présentation
  - Noyaux
  - SVM multi-classe
- Données
  - Individus et Caractéristiques
  - Echantillons: Apprentissage/Test
- Applications
  - Diagnostic cytologique (exemple R)
  - RI sélective (basée sur des prédicteurs de difficulté des requêtes)

# SVM

Séparateur à Vaste Marge

# SVM - Introduction

- Classe d'algorithmes d'apprentissage supervisé
- But:
  - Initial:  discrimination - prévision d'une variable qualitative binaire
  - Généralisé: prévision d'une variable quantitative
- Recherche le hyperplan de marge optimale de séparation (discrimination dichotomique)
- Trouver un classifieur (ou une fonction de discrimination) avec une capacité de généralisation (qualité de prévision) maximale

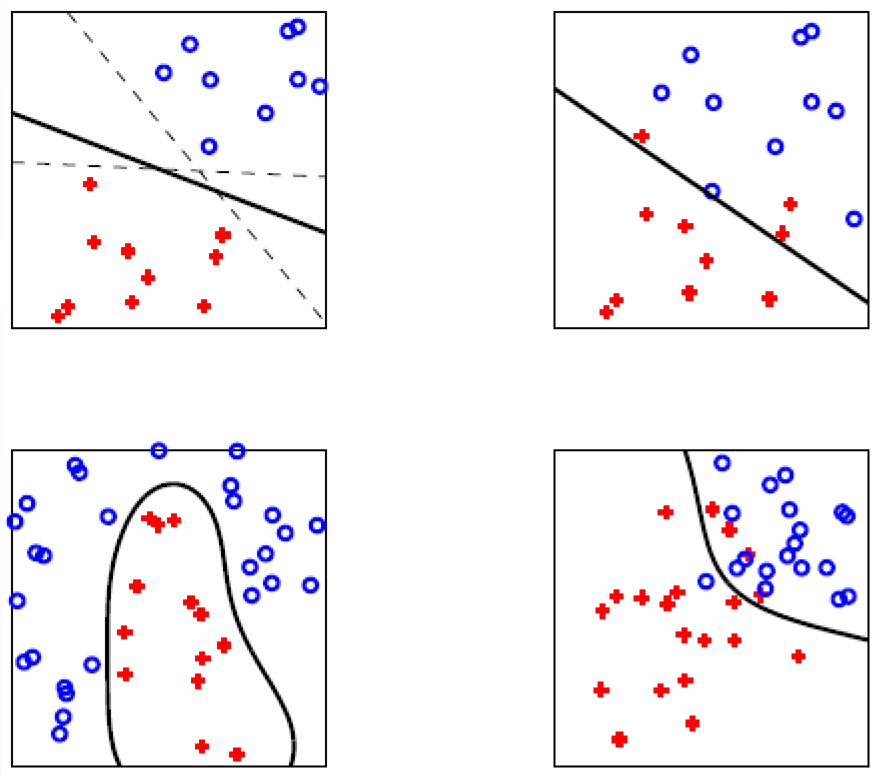
# SVM - Historique

- Marge maximale + Fonction noyau → SVM
- **Idée des hyperplans à marge maximale:**
  - **1963**: V. Vapnik et A. Lerner[1]
  - **1973**: Richard Duda et Peter Hart[2]
- **Idée des fonctions noyaux:**
  - **1909**: théorème de Mercer[3]
  - **1964**: Aizermann et al. [4] – fonctions noyaux dans l'apprentissage artificiel
- Article fondateur: **1992**: Boser et al. [5]
- **1995**: livre de Vapnik[6]
- **1997**: brevet américain

[1] ***(en)*** Vladimir Vapnik et A. Lerner, Pattern Recognition using Generalized Portrait Method, Automation and Remote Control, 1963
[2] ***(en)*** Richard O. Duda, Peter E. Hart, David G. Stork, Pattern classification, Wiley-interscience, 1973
[3] ***(en)*** J. Mercer, Functions of positive and negative type and their connection with the theory of integral equations. Philos. Trans. Roy. Soc. London, A 209:415--446, 1909
[4] M. Aizerman, E. Braverman, and L. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning, Automation and Remote Control 25:821--837, 1964
[5} ***(en)*** Bernhard E. Boser, Isabelle M. Guyon, Vladimir N. Vapnik, A Training Algorithm for Optimal Margin Classifiers [archive] In Fifth Annual Workshop on Computational Learning Theory, pages 144--152, Pittsburgh, ACM. 1992
[6] V. Vapnik, The nature of statistical learning theory, N-Y, Springer-Verlag, 1995

# SVM – Discrimination binaire (I)

- Exemples de quatre types de problèmes de discrimination binaire
  - Il s'agit de séparer les points bleus des croix rouges
  - La frontière de décision est représentée en noir



# SVM – Discrimination binaire (II)

- But: construire une fonction de décision pour associer à chaque observation sa classe
- Cadre probabiliste; observations = vecteurs $\boldsymbol{x} \in \mathbb{R}^p$.
- On suppose l'existence d'une loi inconnue $\mathbb{P}(\boldsymbol{x}, y)$ sur $(\mathbb{R}^p, \{-1, 1\})$
- L'estimateur de la fonction de décision idéale $D : \mathbb{R}^p \to \{-1, 1\}$
  - qui minimise pour toutes les observations la probabilité d'erreur $\mathbb{P}(D(\boldsymbol{x}) \neq y \mid \boldsymbol{x})$
- On suppose l'existence d'un échantillon (ensemble d'apprentissage)

$$\{(\boldsymbol{x}_i, y_i),\ i = 1, n\}$$

  - i.i.d. de loi parente $\mathbb{P}(\boldsymbol{x}, y)$

# SVM linéaires – Problème de discrimination linéaire

- Problème de discrimination linéairement séparable:
  - ∃ une fonction de décision linéaire (séparateur linéaire) qui classe correctement toutes les observations de l'ensemble de l'apprentissage
  - $f$ est une fonction caractéristique

$$D(\boldsymbol{x}) = \text{signe}\big(f(\boldsymbol{x})\big) \text{ avec } f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x} + a, \quad \boldsymbol{v} \in \mathbb{R}^p \text{ et } a \in \mathbb{R}$$

- On peut associer une frontière de décision à cette fonction de décision linéaire

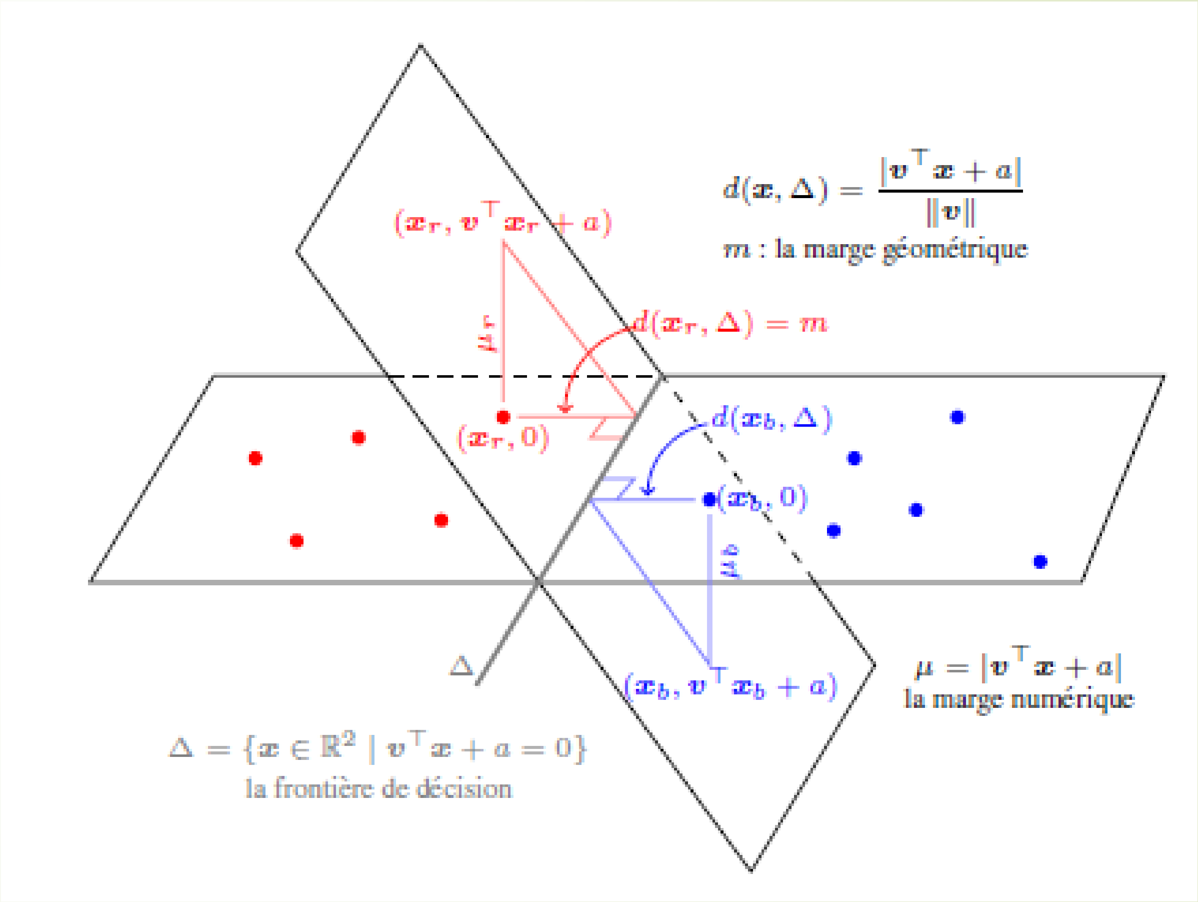
$$\Delta(\boldsymbol{v}, a) = \{\boldsymbol{x} \in \mathbb{R}^p \mid \boldsymbol{v}^\top \boldsymbol{x} + a = 0\}$$

# SVM linéaire – La marge d'un classifieur

- Deux marges pour un même classifieur linéaire:
- Géométrique: la plus petite distance d'un point de l'échantillon à la frontière de décision
- Numérique: la plus petite valeur de la fonction de décision atteinte sur un point de l'echantillon

# SVM linéaire – La marge d’un classifieur

- Illustration de deux notion de marge sur un exemple de discrimination linéaire séparable (2 dimensions)



# SVM linéaire – Maximisation de la marge d'un classifieur

- Maximiser la marge → maximiser la confiance → minimiser la probabilité d'erreur associée au classifieur:

$$\max_{\boldsymbol{v},a} \underbrace{\min_{i\in[1,n]} \operatorname{dist}(\boldsymbol{x}_i, \Delta(\boldsymbol{v}, a))}_{\text{marge : } m}$$

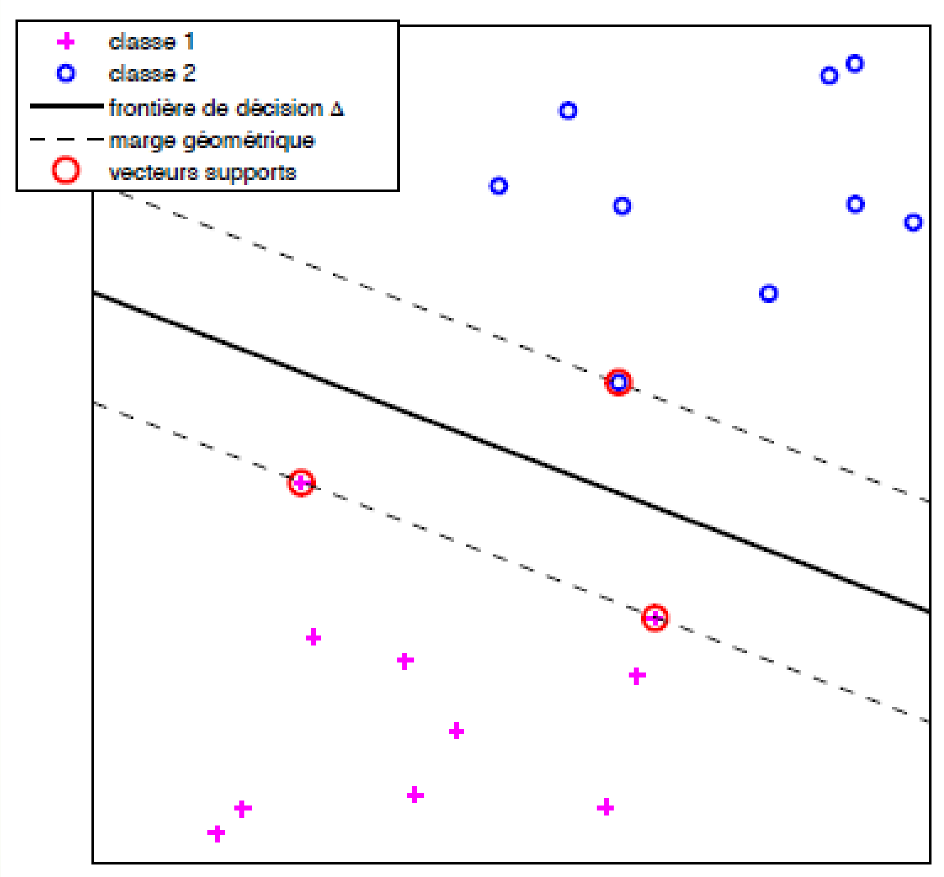
- Définition: SVM sur des données linéairement séparables:

*Soit* $\{(\boldsymbol{x}_i, y_i);\ i = 1, n\}$ *un ensemble de vecteurs formes étiquetées avec* $\boldsymbol{x}_i \in \mathbb{R}^p$ *et* $y_i \in \{1, -1\}$. *Un séparateur à vaste marge linéaire (SVM et support vector machine) est un discriminateur linéaire de la forme :* $D(x) = \mathit{signe}\big(\boldsymbol{w}^\top \boldsymbol{x} + b\big)$ *où* $\boldsymbol{w} \in \mathbb{R}^p$ *et* $b \in \mathbb{R}$ *sont donnés par la résolution du problème suivant :*

$$\begin{cases} \min\limits_{\boldsymbol{w},b} & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{avec} & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \qquad i = 1, n \end{cases}$$

# SVM linéaire – Vecteurs support

- Les vecteurs support definissent la frontière de décision



# SVM linéaire – Le cas des données non séparables (I)

- Notions d'écart:

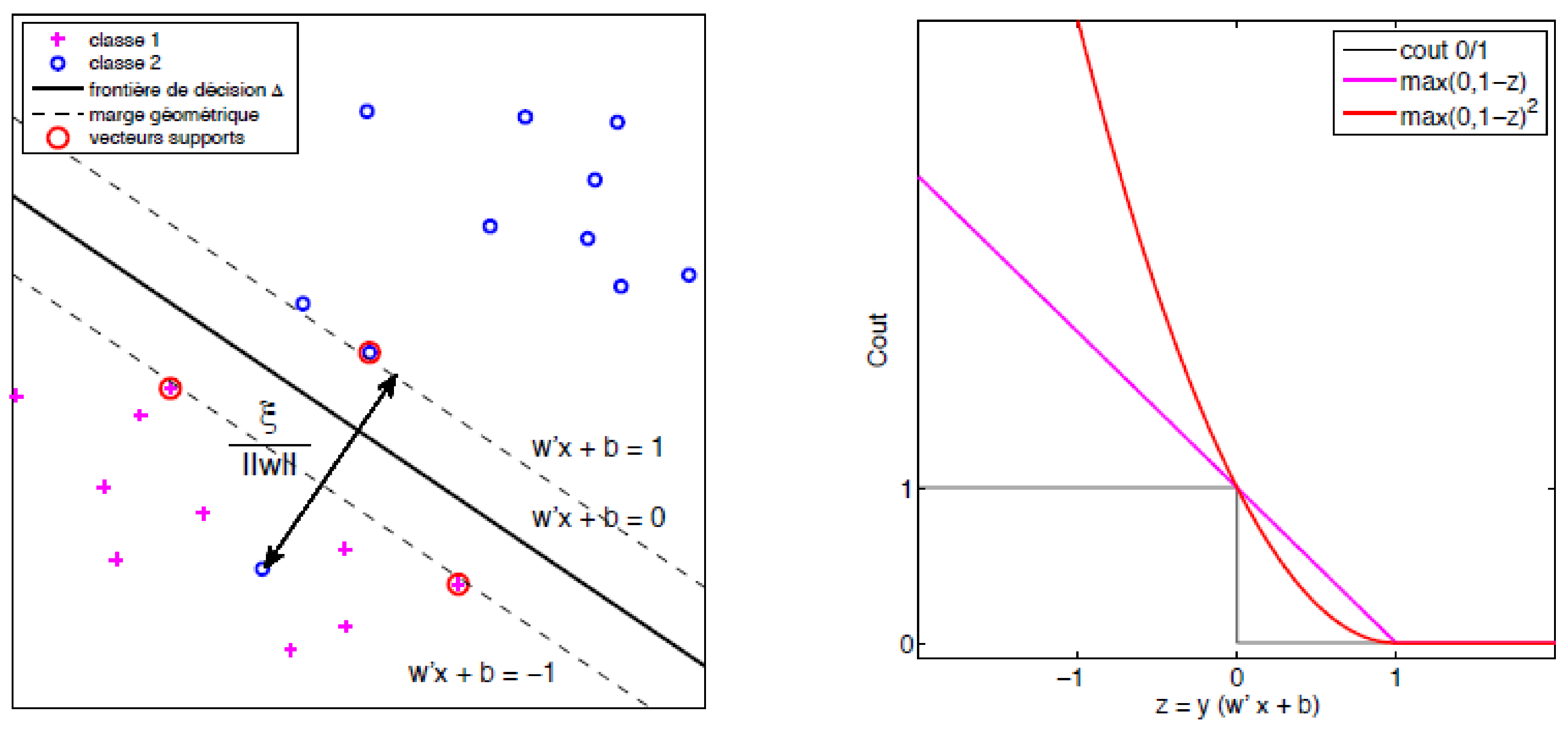
$$\begin{array}{lll} \text{pas d'erreur :} & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \Rightarrow & \xi_i = 0 \\ \text{erreur :} & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) < 1 \Rightarrow & \xi_i = 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0 \end{array}$$

- Notion de « cout charnière »:

$$\xi_i = \max\big(0, 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\big)$$

# SVM linéaire – Le cas des données non séparables (II)

- Notions d'écart et de cout charnière; point bleu mal classé



# SVM – Le cas non linéaire: les noyaux

- Idée: reconsidérer le problème dans un espace de dimension supérieure, éventuellement de dimension infinie
- Dans le nouvel espace: séparation linéaire probable
- On applique aux vecteurs d'entrée $x$ une transformation non-linéaire $\phi$
- L'espace d'arrivée $\phi(X)$ est appelé **espace de redescription**
- L'intérêt de la fonction noyau est double :
  - Le calcul se fait dans l'espace d'origine, ceci est beaucoup moins coûteux qu'un produit scalaire en grande dimension.
  - La transformation $\phi$ n'a pas besoin d'être connue explicitement, seule la fonction noyau intervient dans les calculs → on peut envisager des transformations complexes, et même des espaces de redescription de dimension infinie

# SVM – Le cas non linéaire: les noyaux

- Fonctions noyau:
  - Gaussien: $\mathrm{k}(x, x') = e^{-\gamma\|x-x\prime\|^2}$
  - Sigmoïde: $\mathrm{k}(x, x') = tanh(\gamma x^T x')$
  - Polynomial: $\mathrm{k}(x, x') = (\gamma x^T x')^{\ d}$
- $d$ représente le degré polynomial
- $\gamma$ représente un paramètre d'ajustement

# SVM multi-class

- SVM peuvent être adaptés pour traiter les problèmes multi-classe
- Idée: transformer le problème à « k » classes en « n » classifieurs binaires
- Le classement est donné par le classifieur qui répond le mieux
- !!! Beaucoup d'exemples négatifs !!!
- Stratégies:
  - 1 contre 1: chaque classe est comparée à chaque classe; classement donné par le vote majoritaire ou un graphe acyclique de décision
  - 1 contre toutes: la classe en cause est étiquetée 1, toutes les autres sont étiquetées -1; le classifieur avec la plus élevée valeur de confiance est élu

# SVM – Exemples d'applications

- Classification de données biologiques/physiques
- Classification de documents numériques
- Classification d'expressions faciales
- Classification de textures
- E-learning
- Détection d'intrusion
- Reconnaissance de la parole
- CBIR : Content Based Image Retrieval

# Données

Jeux de données pour apprentissage

# Données: Individus et Caractéristiques

- Exemple de jeu de données
- Lignes = individus (observations)
- Colonnes = caractéristiques
  - Caractéristiques qualitatives
  - Caractéristiques quantitatives
- Une colonne avec la classe correspondante

# Données: Echantillons Apprentissage/Test

- Contexte de l'apprentissage supervisé
- Découpage du jeu de données: apprentissage/test
  - Validation croisée (10-fold cross-validation)
  - Méthode d'exclusion (« leave-one-out »)

# Application 1

Diagnostic cytologique (exemple R)

# Diagnostic cytologique (exemple R)

- Les données issues d'une base dévolue à la comparaison des techniques de modélisation et d'apprentissage:
  - http://archive.ics.uci.edu/ml/
- Données concernant le cancer du sein: nature maligne ou bénigne de la tumeur à partir des variables biologiques
- `Préparation données`
- `Extraction des échantillons`
- `SVM: options par défaut`
- `Options par cross-validation`
- `Noyaux`
  - `Défaut: gaussien`
  - `Polynomial`
  - `Sigmoïdal`
- `Echantillon de test`

# Application 2

[Chifu & Mothe, 2014]: Expansion sélective de requêtes par apprentissage