Apprentissage, réseaux de neurones et modèles graphiques (RCP209)

Machines à vecteurs de support
Support Vector Machines (SVM)

Marin FERECATU & Michel Crucianu
(prenom.nom@cnam.fr)

http://cedric.cnam.fr/vertigo/Cours/ml2/

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

Plan du cours

Objectif

"La raison d'être des statistiques, c'est de vous donner raison." — Abe Burrows

Machines à vecteurs de support (Support Vector Machines SVM) et méthodes à noyau :

- Separateurs à vaste marge
- Cas linéairement séparable
- Cas non-séparable linéairement
- Astuce à noyau
- SVM non linéaire

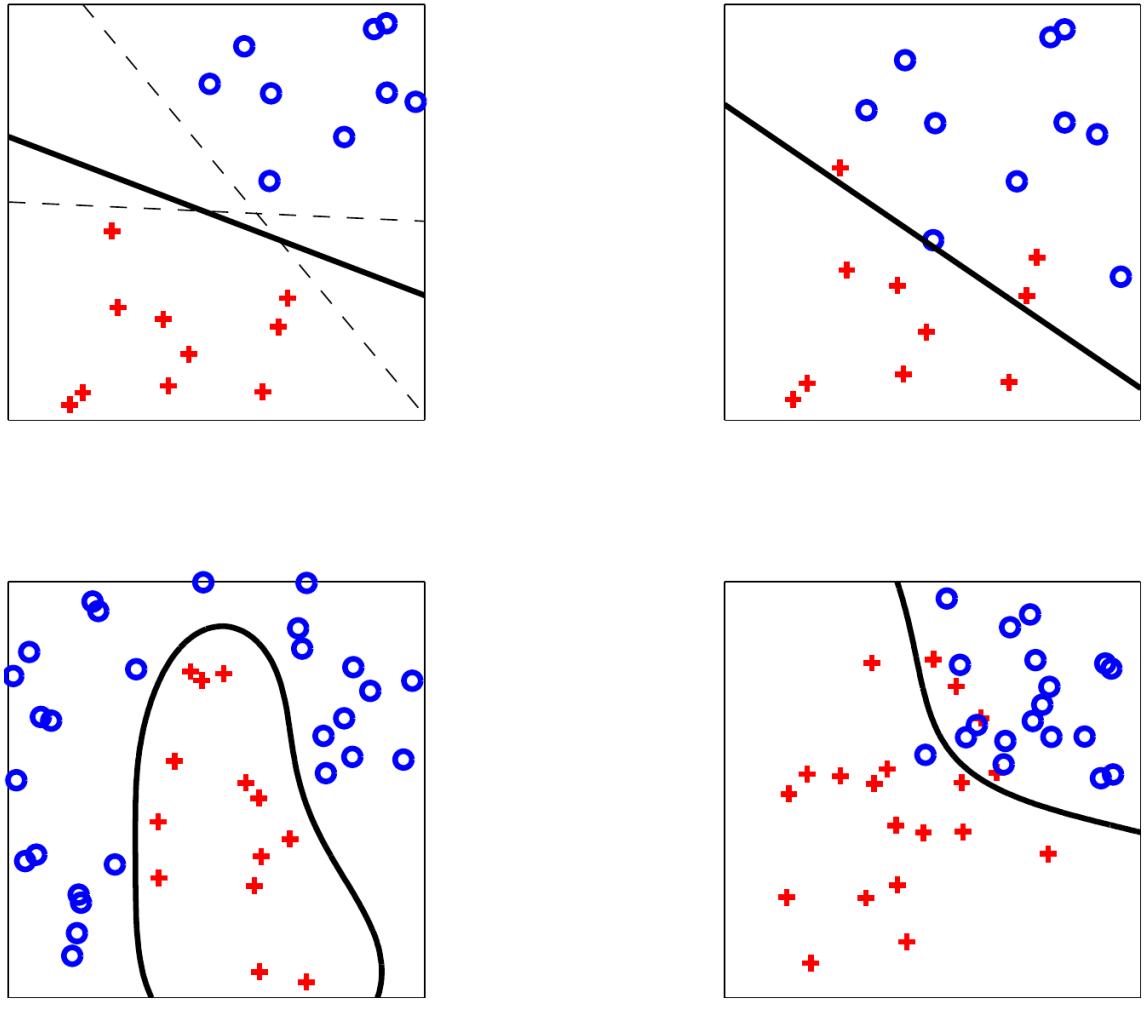
Objectif

SVM et méthodes à noyau :

- SVM pour la régression
- One-class SVM
- Principe des méthodes à noyaux
- Kernel PCA, Kernel CCA
- SVM à noyaux multiples (Multiple Kernel Learning - MKL)
- Noyaux pour des données structurés
- Applications

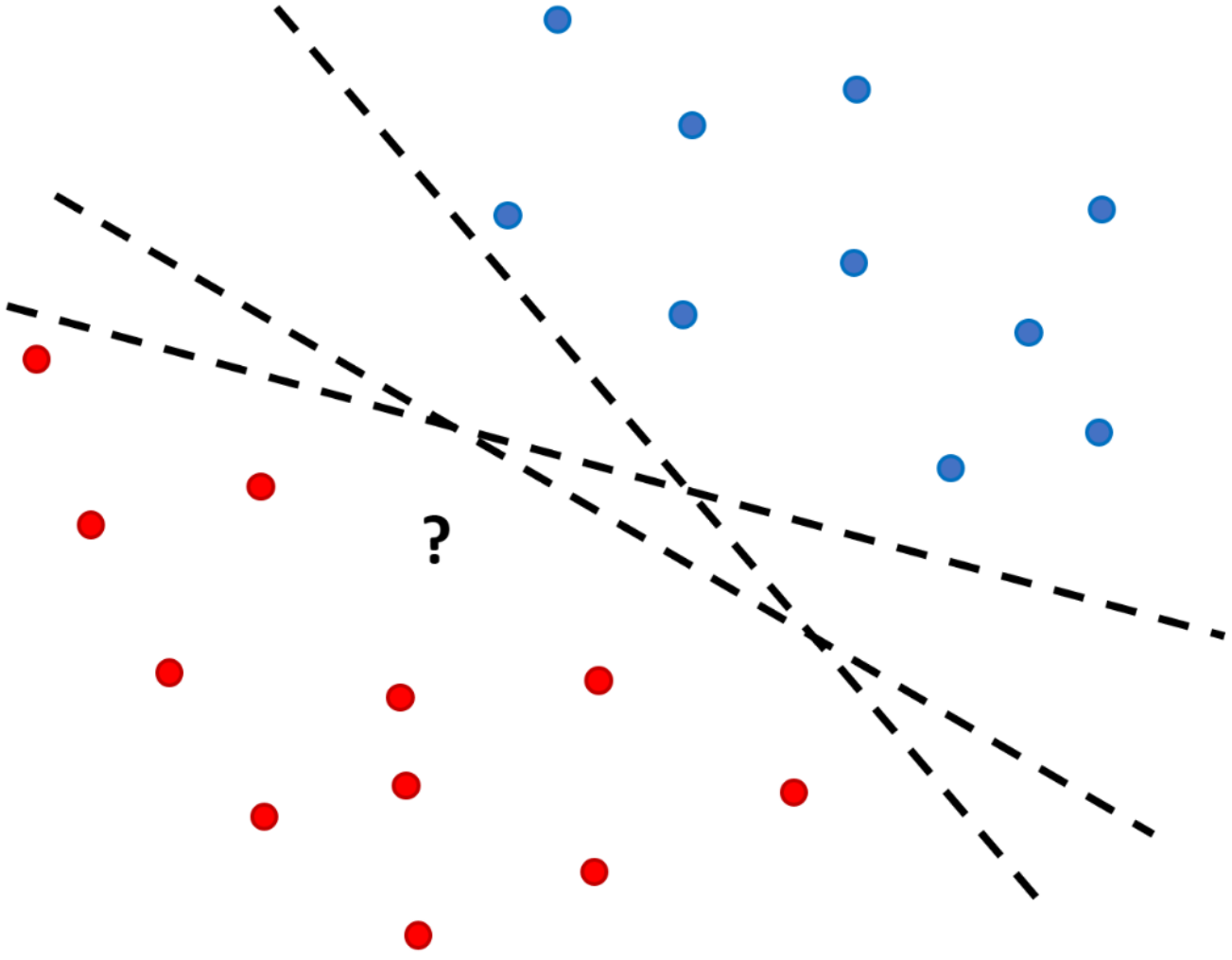
Plan du cours

Problèmes de classification :



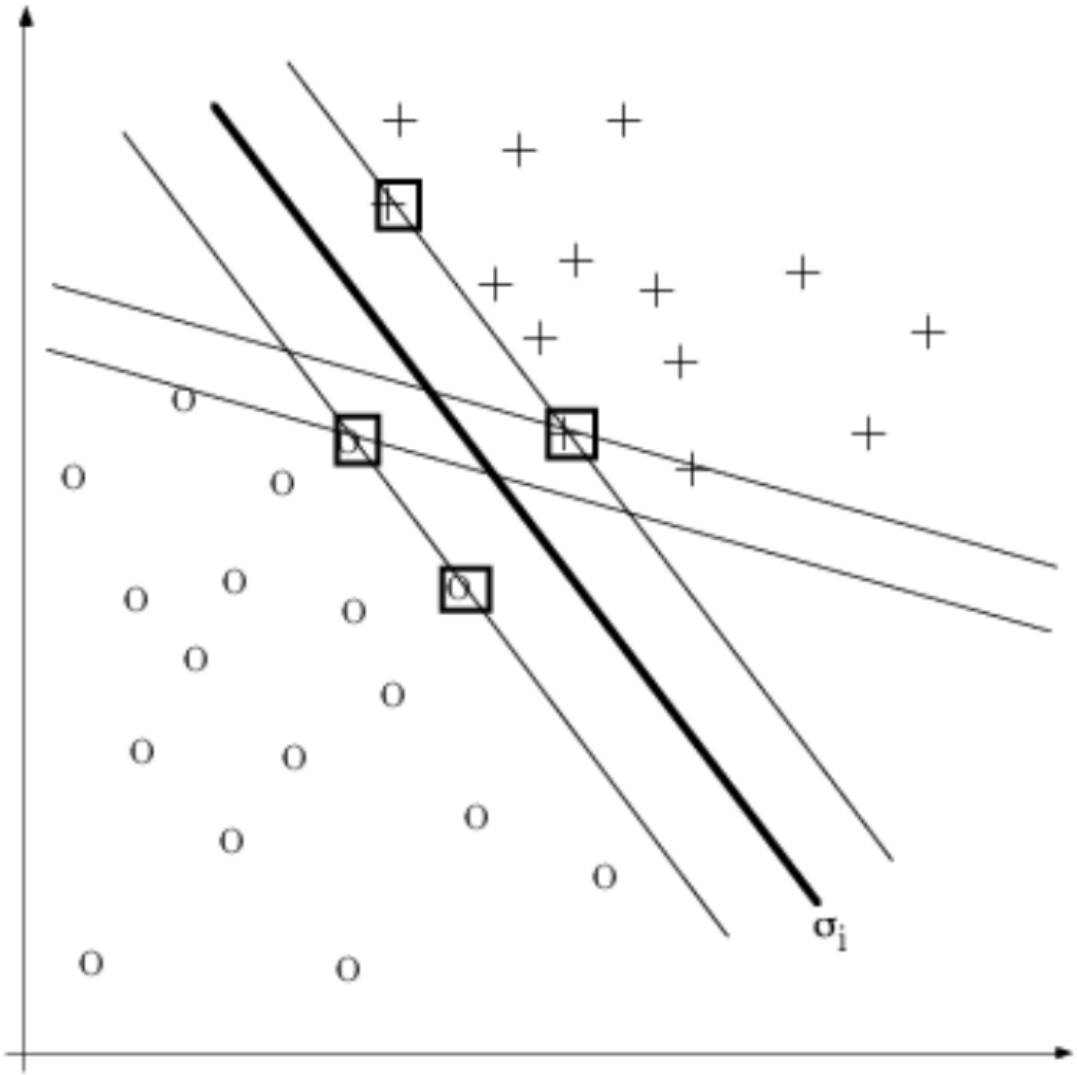
Linéaire (haut) vs. non-linéaire (bas).

Séparation linéaire



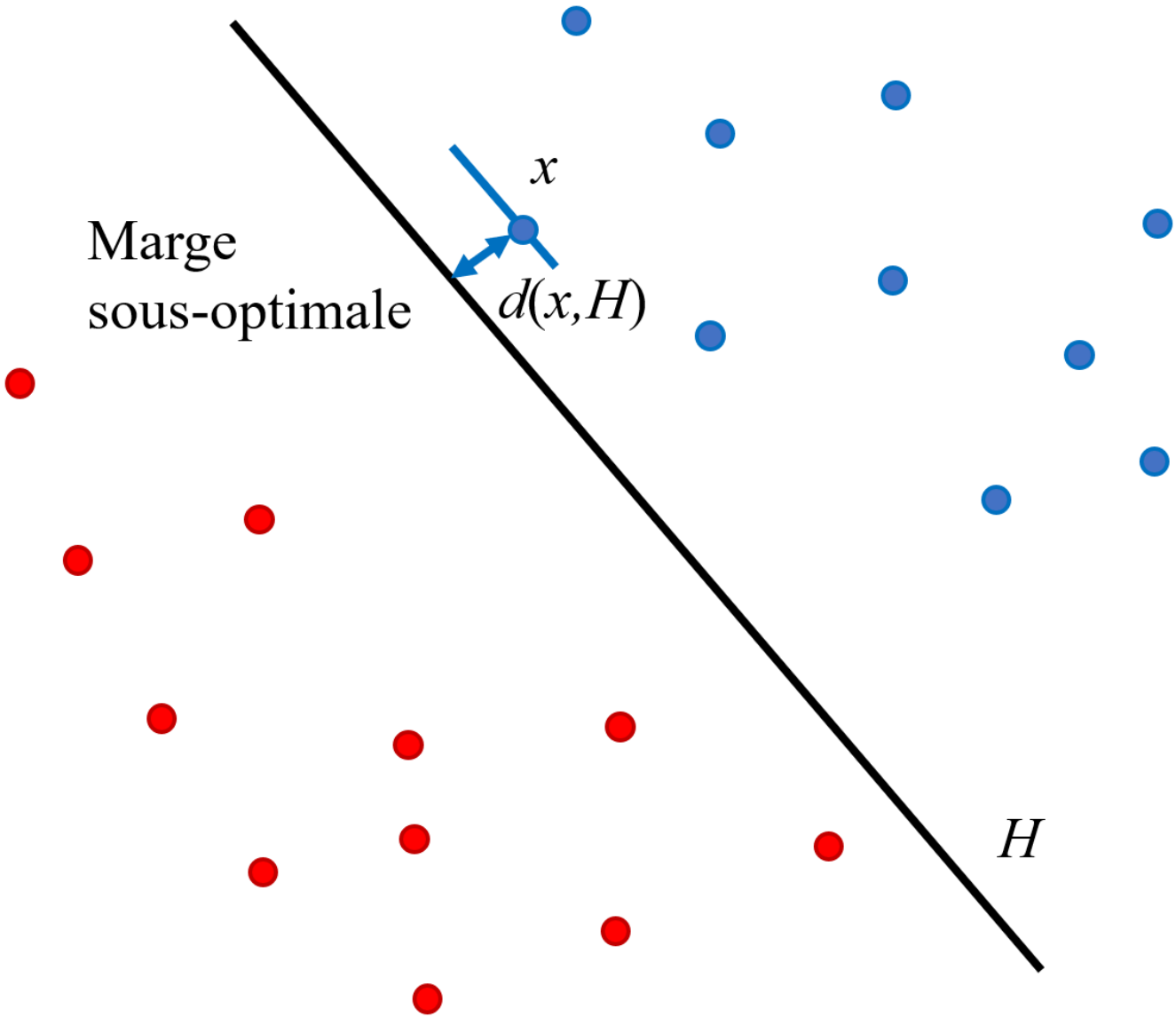
Séparateurs linéaires.

Séparation linéaire et marge



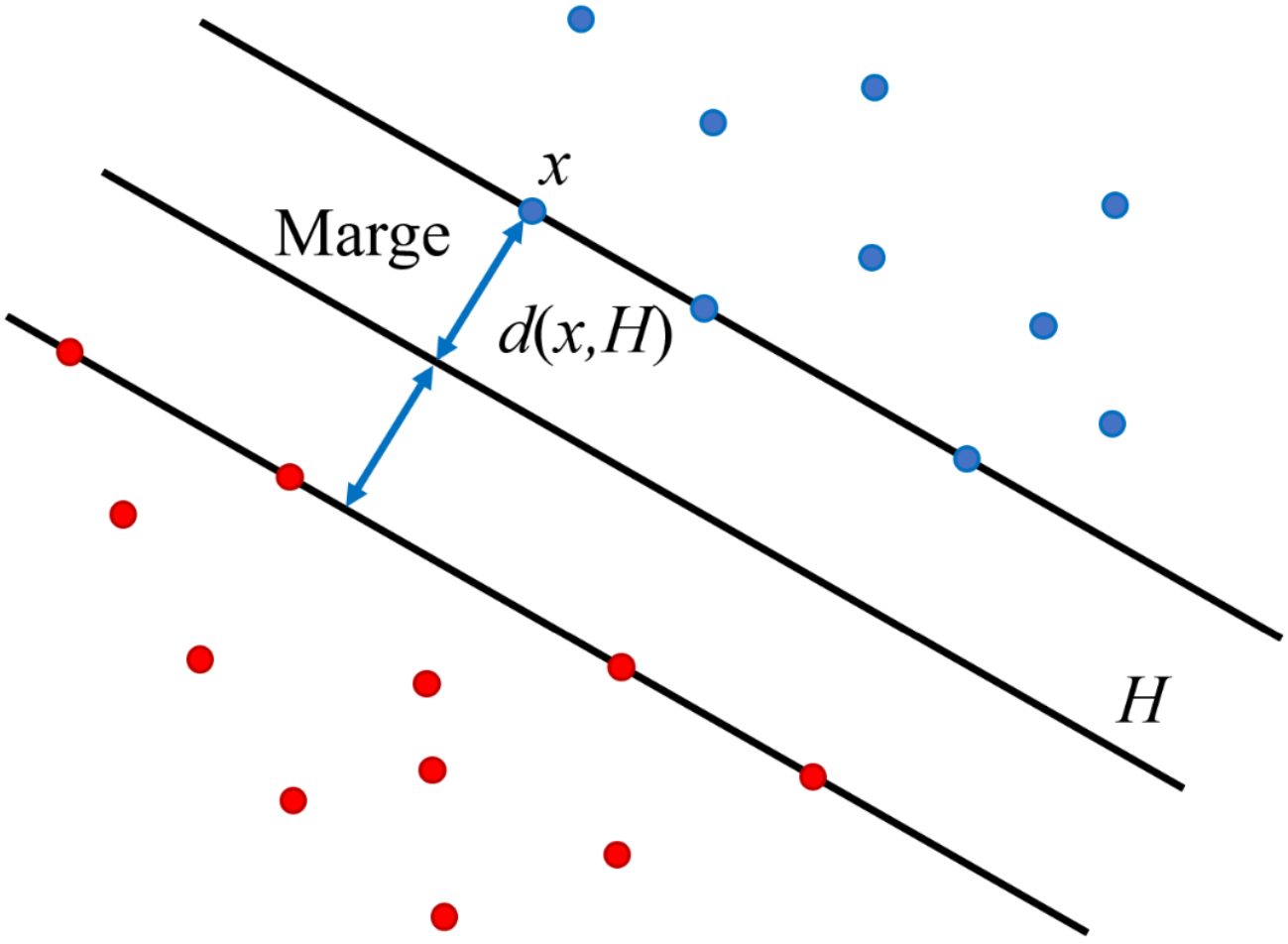
Marge des séparateurs linéaires.

Séparation linéaire et marge



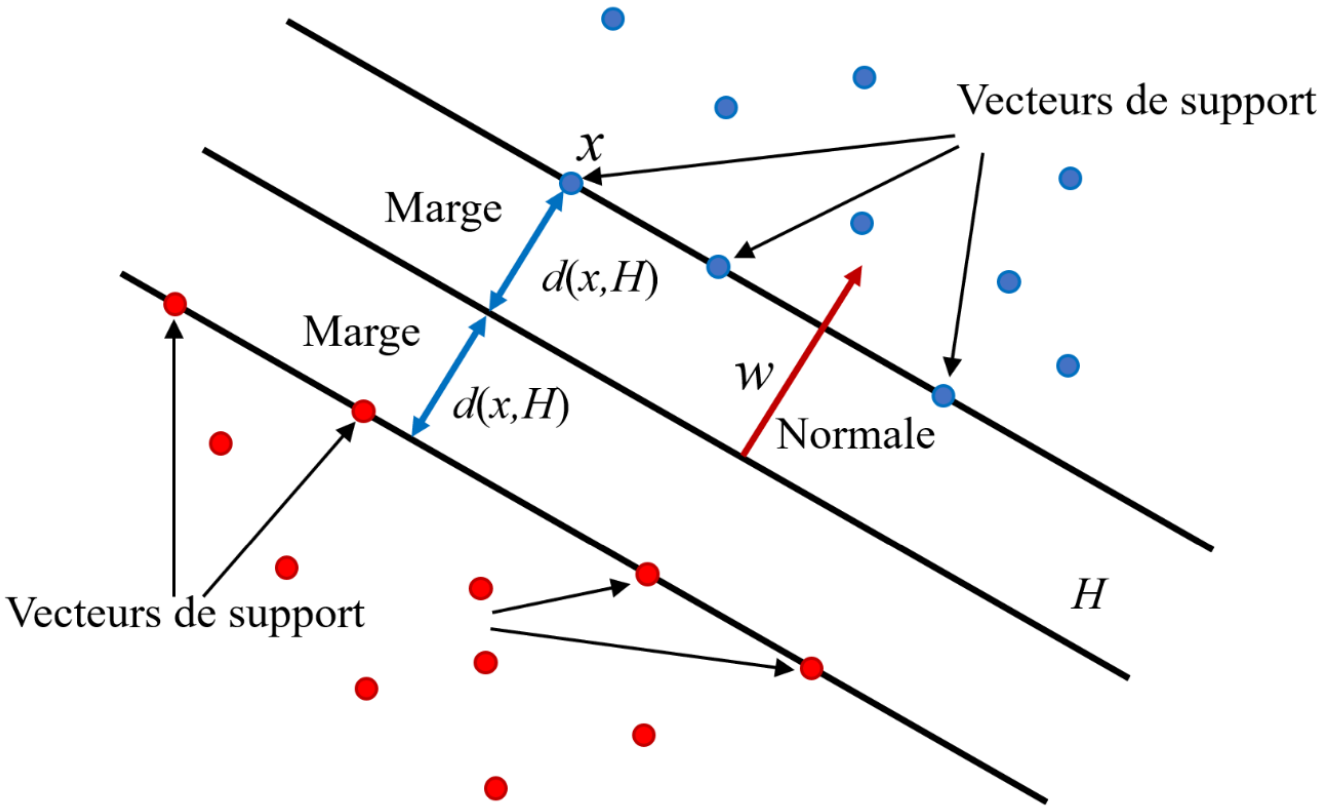
Marge des séparateurs linéaires.

Séparation linéaire et marge



Marge des séparateurs linéaires.

Séparation linéaire et marge

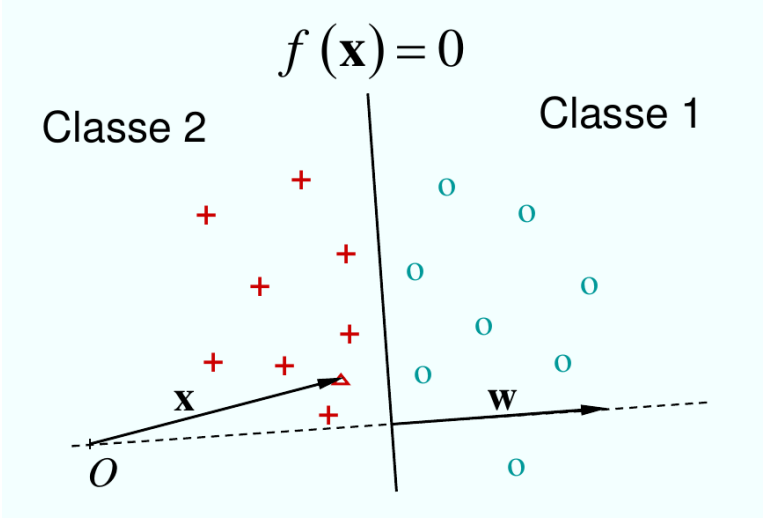


Marge des séparateurs linéaires.

Séparation linéaire et marge

Marge : Distance entre le plus proche exemple d'apprentissage et la surface de séparation.

Bas d'apprentissage : $\{(x_i, y_i), i = 1, \ldots, n\}$, $x_i \in R^d$, $y_i \in \{-1, 1\}$



Fonction de décision : $f(x) = w^T x + b = 0$

- $f(x) = 0$: hyperplan (surface) de séparation
- $f(x) > 0$: classe 1 ($y_i = 1$)
- $f(x) < 0$: classe 2 ($y_i = -1$)

Séparation linéaire et marge

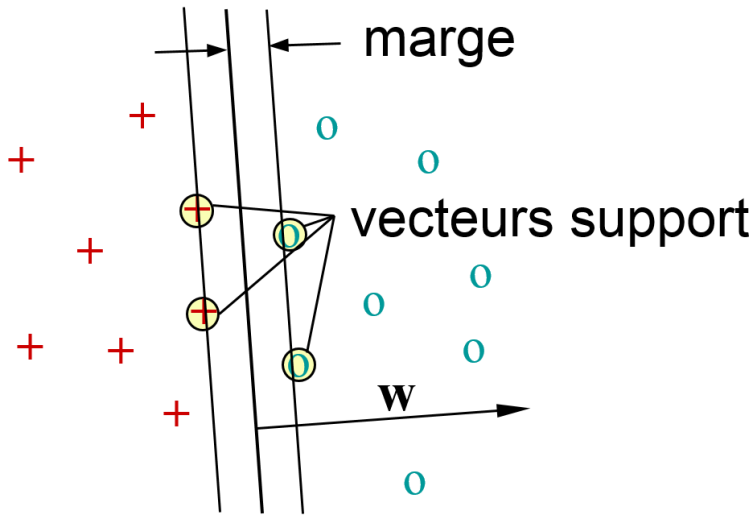
Fonction de décision : $f(x) = w^T x + b = 0$

Paramètres :

- w est la normale à l'hyperplan,
- b est le décalage par rapport à l'origine
- Les paramètres w et b ne sont pas uniques. kw et kb donnent la même surface de séparation : $kw^T x + kb = k(w^T x + b) = 0$

Séparation linéaire et marge

Quelle fonction de décision choisir : $f(x) = w^T x + b = 0$



Solution : celle qui maximise la marge.

Séparation linéaire et marge

Si x_s est un support vecteur, et $H = \{x | w^T x + b = 0\}$ alors la marge est :

$$\text{marge} = d(x, H) = \frac{|w^T x_s + b|}{||w||}$$

On impose la *condition de normalisation* $|w^T x_s + b| = 1$ pour les vecteurs de support x_s :

$$\text{marge} = \frac{1}{||w||}$$

Plan du cours

SVM linéaire (cas séparable)

Optimisation de la marge : optimisation sous contraintes (problème primal)

$$\begin{cases} \min_{w,b} \frac{1}{2}||w||^2 \\ t.q.\ y_i(w \cdot x_i + b) \geq 1, i = 1, \ldots, n \end{cases}$$

La résolution de ce problème peut se faire directement (méthodes stochastique de type Gauss-Seidel, algorithmes de point intérieur, de type Newton ou de type gradient conjugué)
Il est toutefois mieux de passer à la formation duale de ce problème :

- Le dual est un problème quadratique de taille n (égal au nombre d'observations)
- Pour ce type de problèmes (optimisation quadratique) il existe des algorithmes bien étudiés et très performants
- La formulation duale fait apparaître la matrice de Gram XX^T ce qui permet de gérer le cas non linéaire à travers des noyaux.

SVM linéaire (cas séparable)

On introduit les multiplicateurs α de Lagrange :

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w^T x_i + b) - 1]$$

Les conditions nécessaires d'optimum :

$$\frac{\partial L}{\partial b} L(w, b, \alpha) = 0 \Longrightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial w} L(w, b, \alpha) = 0 \Longrightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

SVM linéaire (cas séparable)

Par substitution on obtient le problème dual :

$$\begin{cases} \max\limits_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ t.q. \\ \alpha_i \geq 0, i = 1, \ldots, n \ \text{(admissibilité duale)} \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \ \text{(stationnarité)} \end{cases}$$

- Les vecteurs de support sont ceux pour lesquels $\alpha_i > 0$
- Ajouter des échantillons à l'ensemble d'apprentissage qui ne sont pas des vecteurs supports n'a aucune influence sur la solution finale
- b est obtenu à partir de la relation $|x_s^T w + b| = 1$ valable pour tous les vecteurs de support

SVM linéaire (cas séparable)

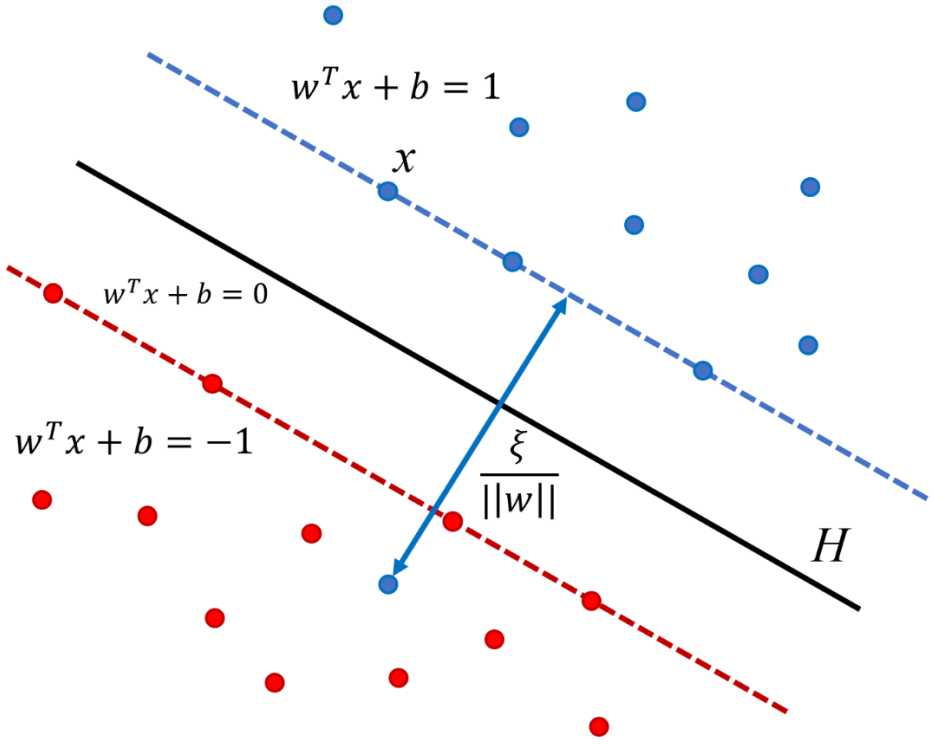
La fonction de décision permettant de classer une nouvelle observation x est

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i x_i^T x + b$$

- L'hyperplan solution ne dépend que du produit scalaire entre le vecteur d'entrée et les vecteurs de supports. Cette particularité est l'origine de la 2eme innovation majeure des SVM : le passage par un espace de description grâce à des fonctions noyau.

Plan du cours

SVM linéaire (cas non séparable)



Dans le cas ou les données ne sont pas séparables linéairement on utilise une technique dite de **marge souple**, qui tolère les mauvais classements :

- Rajouter des variables de relâchement des contraintes ξ_i
- Pénaliser ces relâchements dans la fonction objectif.

SVM linéaire (cas non séparable)

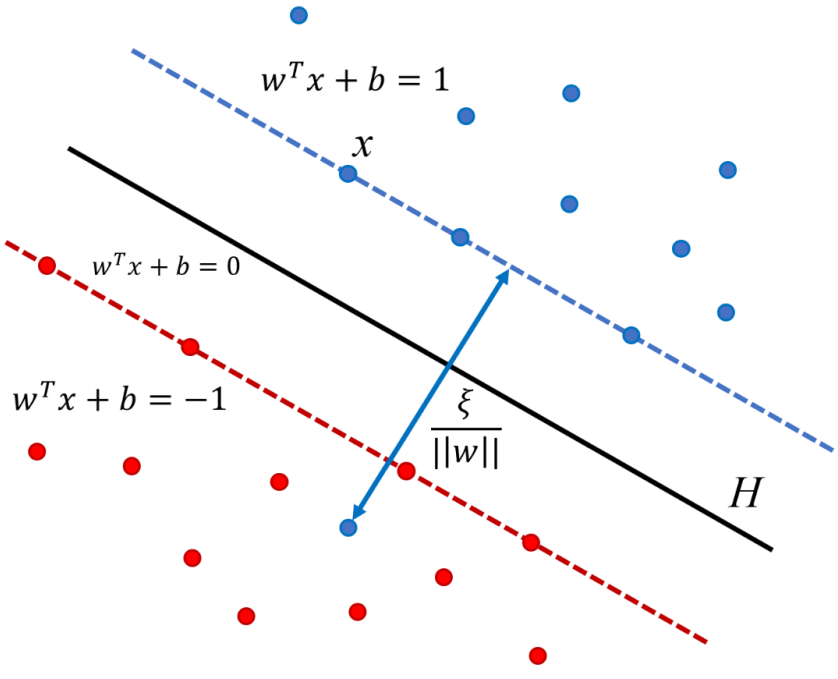
L'idée : modéliser les erreurs potentielles par des variables d'écart positives ξ_i associées aux observations $(x_i, y_i), i = 1, \ldots n$.
Si un point (x_i, y_i) vérifie la contrainte de marge $y_i w^T x_i + b) \geq 1$ alors la variable d'écart (qui est une mesure du cout de l'erreur) est nulle.
Nous avons donc deux situations :

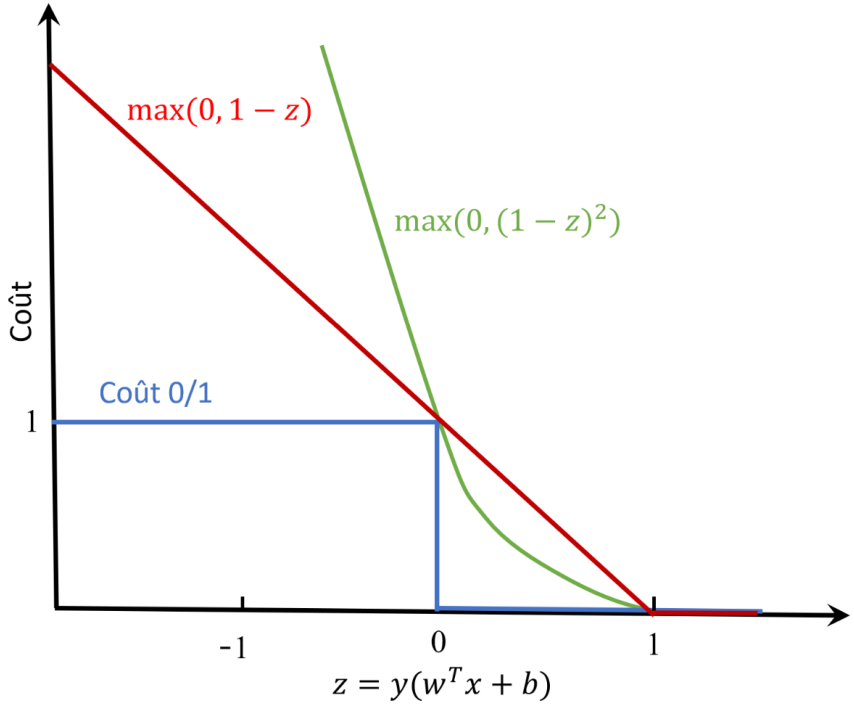
- Pas d'erreur : $y_i(w^T x_i + b) \geq 1 \Longrightarrow \xi_i = 0$
- Erreur : $y_i(w^T x_i + b) < 1 \Longrightarrow \xi_i = 1 - y_i(w^T x_i + b) > 0$

SVM linéaire (cas non séparable)

On associe à cette définition une fonction cout appelée « cout charnière » :

$$\xi_i = \max\left(0, 1 - y_i(w^T x_i + b)\right)$$





Un seul point est mal classé (point bleu). L'écart mesure la distance du point à la marge numérique de l'hyperplan séparateur.

SVM linéaire (cas non séparable)

Problème d'optimisation dans le cas des données non-séparable :

$$\begin{cases} \min\limits_{w,b} \begin{cases} \frac{1}{2}||w||^2 \\ \sum_{i=1}^{n} \xi_i \end{cases} \\ t.q. \\ y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, \ldots, n \\ \xi_i \geq 0, i = 1, \ldots, n \end{cases}$$

- Si toutes les variables d'écart $\xi_i = 0$, on retrouve le problème séparable linéairement
- Puisque il faut minimiser les deux termes simultanément on introduit une variable d'équilibrage $C > 0$ qui permet d'avoir une seule fonction objectif dans le problème d'optimisation :

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

SVM linéaire (cas non séparable)

Problème d'optimisation dans le cas des données non-séparable :

$$\begin{cases} \min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \\ t.q. \\ y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, \ldots, n \\ \xi_i \geq 0, i = 1, \ldots, n \end{cases}$$

- C est une variable de pénalisation des points mal classés faisant un compromis entre la largeur de la marge et les points mal classés.
- ξ_i s'appellent aussi *variables ressort* (anglais : *slack variables*)

SVM linéaire (cas non séparable)

Le problème dual devient :

$$\begin{cases} \max\limits_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ t.q. \\ C \geq \alpha_i \geq 0, i = 1, \ldots, n \ \text{(admissibilité duale)} \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \ \text{(stationarité)} \end{cases}$$

- C joue le rôle d'une constante de régularisation (la régularisation est d'autant plus forte que C est proche de 0 !)
- La différence pour le problème duale entre le cas séparable et non séparable est que les valeurs des α_i sont majorées par C.
- Les points mal classés ou placés dans la marge ont un $\alpha_i = C$
- b est calculé de sorte que $y_i f(x_i) = 1$ pour les points tels que $C > \alpha_i > 0$

La fonction de décision permettant de classer une nouvelle observation x est toujours

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i x_i^T x + b$$

SVM linéaire (cas non séparable)

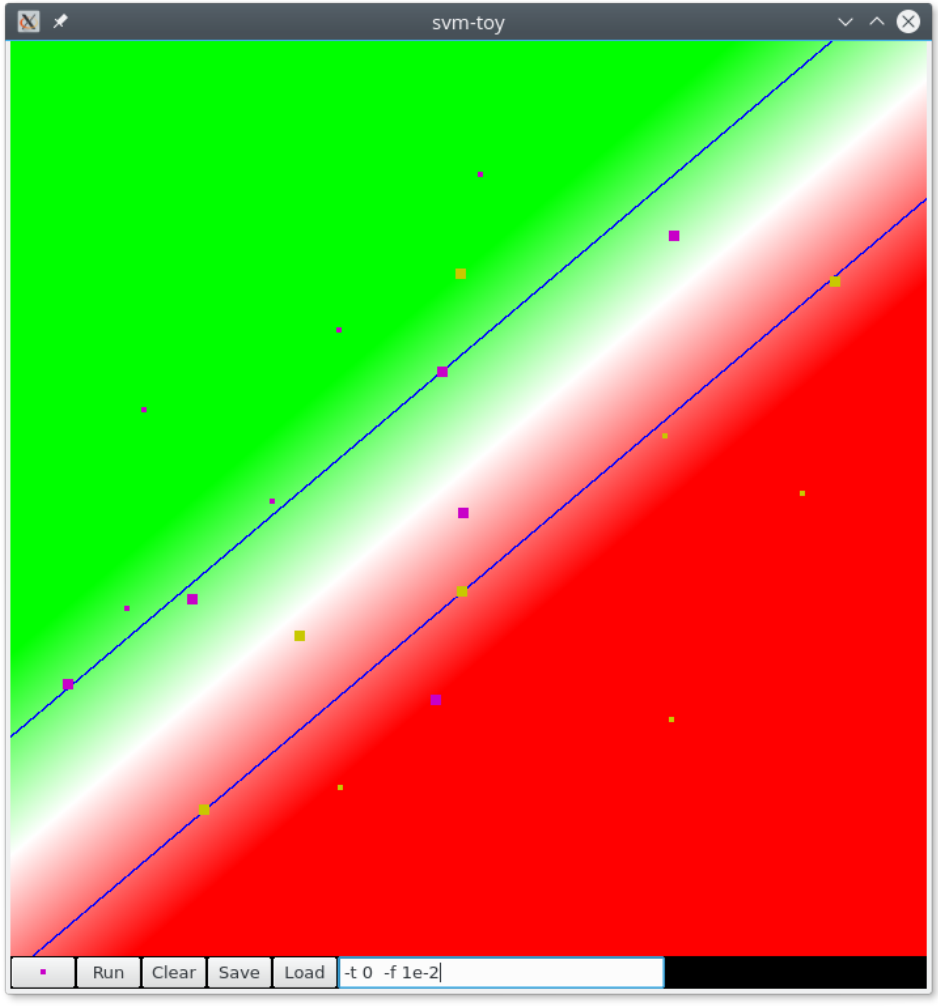
Implémentations software : Torch, LibSVM, LibLinear, Scikit-Learn

- LibSVM, https ://www.csie.ntu.edu.tw/~cjlin/libsvm/
- LibLinear, https ://www.csie.ntu.edu.tw/~cjlin/liblinear/
- Scikit-Learn, http ://scikit-learn.org/

Pratiquement tous les grands environnement de modélisation mathématique possèdent implémentations performantes pour les SVM et méthodes à noyaux (R, Matlab, Mathematica, Scipy, Torch, Scikit-learn, etc.)

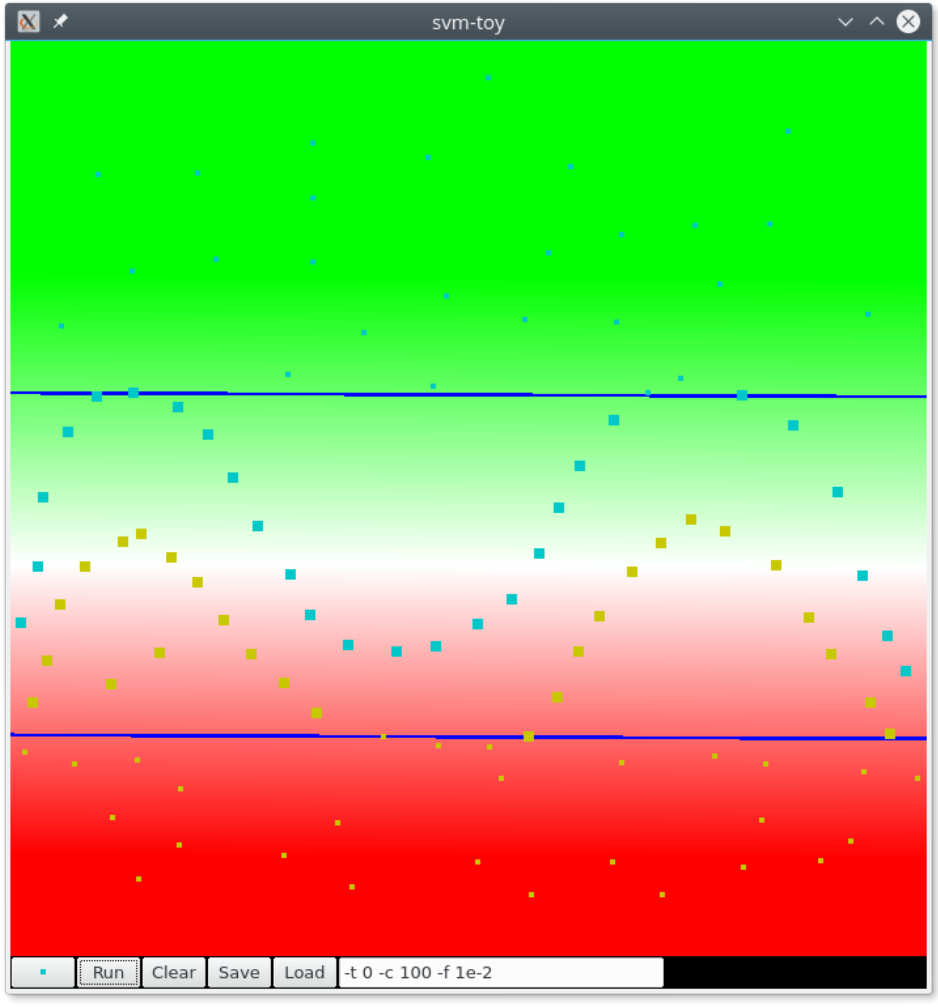
SVM linéaire (cas non séparable)

Séparation linéaire (vecteurs de support en gras) :



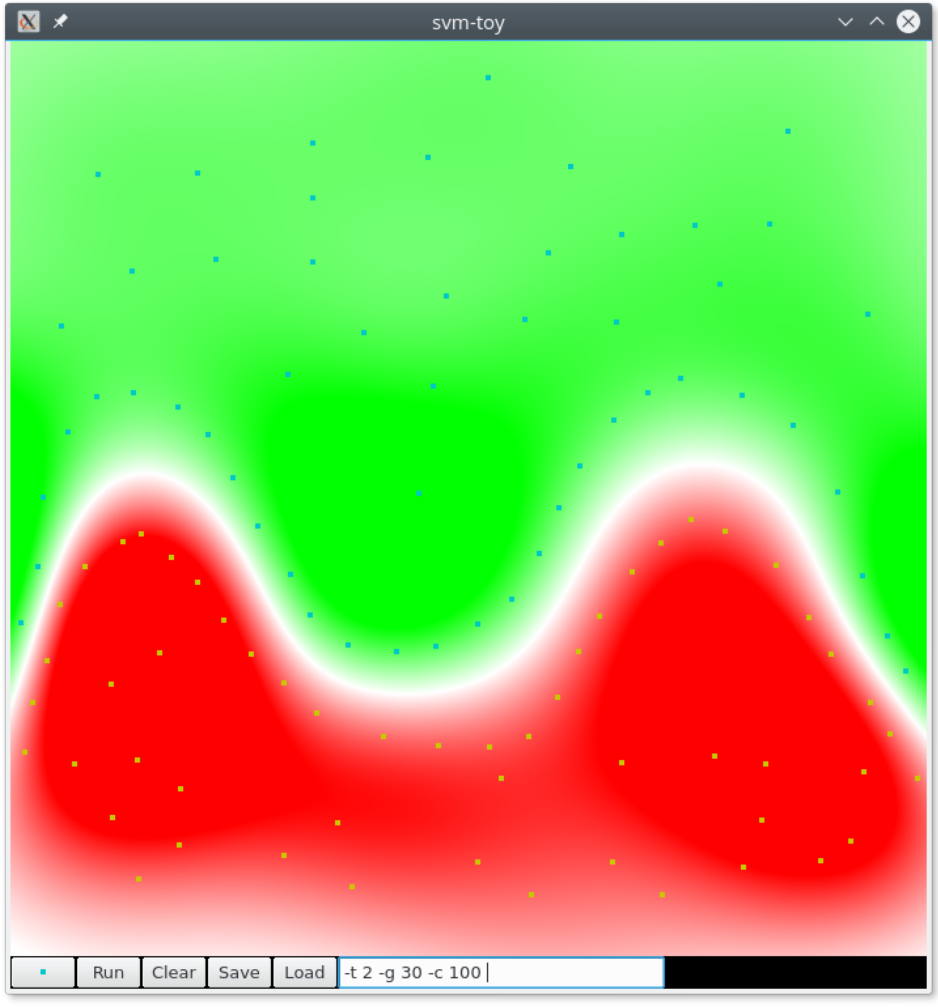
SVM linéaire (cas non séparable)

Séparation linéaire (vecteurs de support en gras) :



SVM linéaire (cas non séparable)

La version à noyaux (séance suivante) permet de séparer mieux les classes :



SVM linéaire (cas non séparable)

Ou même des classes plus compliquées :

Références

Livres, articles, web :

- Steinwart, Christmann, *Support Vector Machines*, Springer 2008
- Scholkopf, Smola, *Learning with Kernels*, The MIT Press, 2001
- Hastie, Tibshirani, Friedman, *The elements of statistical learning : Data mining, inference, and prediction*, New York, Springer Verlag, 2006
- —, *Machines à vecteurs supports (WikiStat)*, http ://wikistat.fr