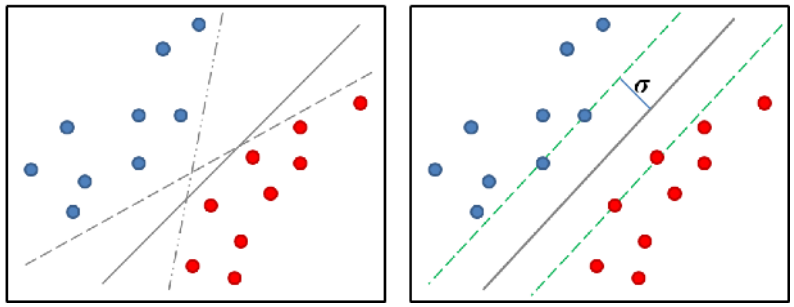
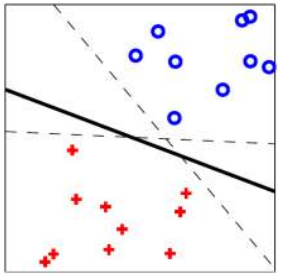
Les Support Vector Machines



Apprentissage statistique : classifieurs linéaires

- On étudie tout d'abord un problème bi-classe :
  - Données étiquetées (+ v.s. −)
  - Trouver un hyperplan qui sépare les données de dimension N
  - Hyperplan : espace de dimension $N-1$
  - En 2d, une droite



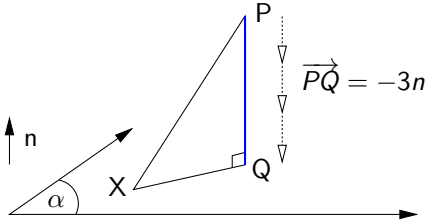
Hypothèse fondamentale

- Données d'apprentissage / de test : distribuées selon la même loi
- En pratique, hypothèse violée à un certain point
- Pour atteindre de bonnes performances, les algorithmes d'apprentissage statistiques doivent disposer d'exemples d'apprentissage suffisamment représentatifs des données de test.

Différentes manières de choisir l'hyperplan car Différents critères à optimiser

Le perceptron

- Distance à l'hyperplan séparateur : $d = \frac{|w'x+b|}{||w||}$



$d(P,\alpha) = |PQ|$. Le point Q est de la forme $P + t\mathbf{n}$, où t est choisi pour que

$$\overrightarrow{XQ}.\mathbf{n} = (\overrightarrow{XP} + \overrightarrow{PQ}).\mathbf{n} = \overrightarrow{XP}.\mathbf{n} + t|\mathbf{n}|^2 = 0.$$

$\Rightarrow$ un seul point et $d(P,\alpha) = |t\mathbf{n}| = \frac{|\overrightarrow{XP}.\mathbf{n}|}{|\mathbf{n}|}$, (valable aussi lorsque $P \in \alpha$).
Si α : $w_1x_1 + \cdots + w_nx_n + b = 0$ (repère orthonormé) :

$$d(P,\alpha) = \frac{|w_1p_1 + \cdots + w_np_n + b|}{\sqrt{w_1^2 + \cdots + w_n^2}}.$$

En effet, on peut prendre $\mathbf{n} = (w_1, \ldots, w_n)^t$ et si $X = (x_1, \ldots, x_n)^t$, on a $\overrightarrow{XP}.\mathbf{n} = w_1(p_1 - x_1) + \cdots + w_n(p_n - x_n) = w_1p_1 + \cdots + w_np_n + b$.

Le perceptron

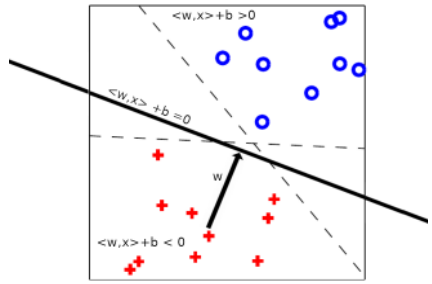
- Distance à l'hyperplan séparateur : $d = \frac{|w'x+b|}{||w||}$
- Critère d'optimisation : minimiser le nombre d'exemples mal classés

Données mal classées :

- Exemples − : $<w,x> + b < 0$
- Exemples + : $<w,x> + b > 0$

Perceptron :

$$min(-\sum_{\text{mal classés}} y_i(w.x_i + b))$$



Gradient

$$\frac{\partial}{\partial w} = -\sum_{\text{mal classés}} y_i x_i$$

$$\frac{\partial}{\partial b} = -\sum_{\text{mal classés}} y_i$$

Apprentissage

- Initialisation aléatoire des poids
- Mise à jour itérative de l'hyperplan : descente gradient (ρ coeff d'apprentissage)
  - $w \leftarrow w + \rho \sum y_i x_i$
  - $b \leftarrow b + \rho \sum y_i$
- Ou descente gradient stochastique (exemple par exemple)

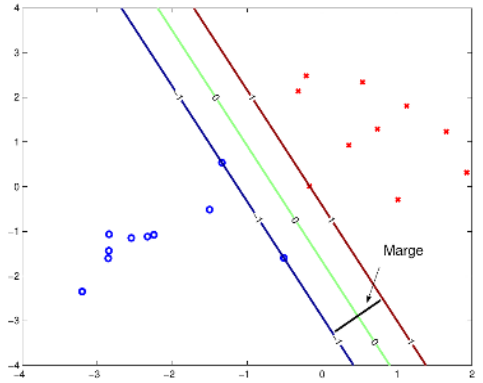
$$\begin{pmatrix} w \\ b \end{pmatrix}_{n-1} + \rho \begin{pmatrix} y_i . x_i \\ y_i \end{pmatrix} \rightarrow \begin{pmatrix} w \\ b \end{pmatrix}_n$$

Conclusion

- Supposons les données linéairement séparables : convergence
  Pas de convergence sinon
- Solutions multiples au problème dans le cas séparable
  Dépendantes de l'initialisation

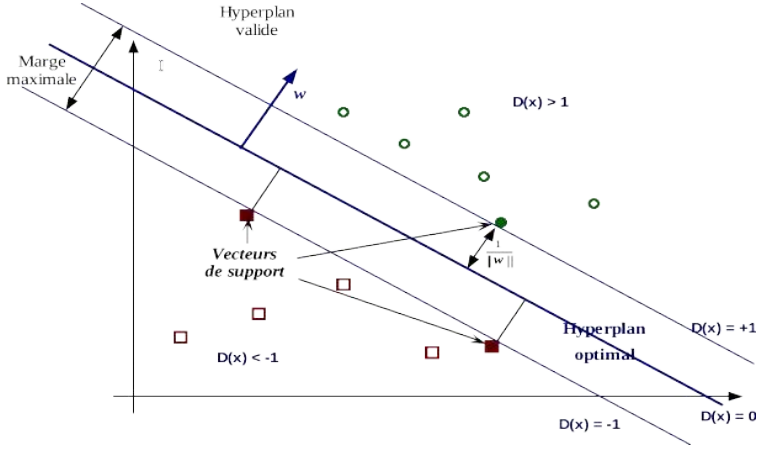
Support Vector Machines (SVM)

- Retour sur les séparateurs linéaires
  - Solutions multiples au problème dans le cas séparable
  - Support Vector Machines (SVM) :
    Frontière avec « no man's land » maximal, hyperplan « épais »



SVM : maximisation de la marge

- Maximise la "marge" ou rayon du corridor :
  distance du point le plus proche à l'hyperplan : $1/||w||$



SVM : maximisation de la marge

- Maximiser la "marge" : assure de bonnes propriétés de généralisation

$$R < R_{emp} + \sqrt{\frac{1}{n}(h(ln(2n/h)+1) - ln(\alpha/4))} \qquad \left(\begin{array}{c} \text{Structural} \\ \text{Risk Minimization} \end{array} \right)$$

  - n : nombre d'exemples d'apprentissage
  - h : VC dimension ($d+1$ pour hyperplans dans $\mathbb{R}^d$)
  - Borne valable avec la proba $1-\alpha$
- Borne sur le risque réel (généralisation)
- Une des raisons principales du succès des SVM
- Maximise la "marge" ou rayon du corridor : distance du point le plus proche à l'hyperplan : $1/||w||$
  Sous la contrainte que tous les points soient bien classés

$$\begin{cases} min & \frac{1}{2}||w||^2 \\ \forall i & y_i(w.x_i + b) \geq 1 \end{cases}$$

dimension de Vapnik-Chervonenkis

Pourquoi ? dans le cas séparable

- Séparable : $\exists$ hyperplan qui sépare les données sans erreur.
  On cherche alors l'hyperplan de marge maximale d'équation :
  $$f(x) \quad = \quad <w,x> +b \quad = \quad w^t.x + b$$
- Calculons pour X_s un vecteur de support, la distance de X_s à H. Soit P le projeté de X_s sur H et A :
  $\overrightarrow{OP} = \overrightarrow{OX_s} + t.\overrightarrow{w}$ où t est choisi telle que $P \in H$:

$$\overrightarrow{w}.\overrightarrow{OP} + b = 0 = \overrightarrow{w}.(\overrightarrow{OX_s} + \overrightarrow{X_sP}) + b = \overrightarrow{w}.(x_s + t.\overrightarrow{w}) + b = 0.$$

  Autrement dit, on a : $(\overrightarrow{w}.x_s + b) + t.||\overrightarrow{w}||^2 = 0$.
- La marge (plus exactement le double de la marge) est alors donnée par :
  $$Marge \quad = \quad 2d(x,H) \quad = \quad 2\frac{|w^t.x_s + b|}{||w||}$$
- w et b ne sont pas uniques, kw et kb donnent le même hyperplan :

$$kw^t.x + kb \quad = \quad k(w^t.x + b) \quad = \quad 0$$

$\Rightarrow$ Condition de normalisation $|w^t.x_s + b| = 1$ pour les vecteurs de support x_s

ce qui conduit à : $\quad Marge = \frac{2}{||w||}$

- On arrive donc au problème d'optimisation (appelé problème primal) :
  $$\begin{cases} \min_{w,b} \frac{1}{2}||w||^2 \\ \text{tel que } y_i \times (w.x_i + b) \geq 1, i = 1, ..., n \end{cases}$$

Relaxation lagrangienne

- Considérons le problème d'optimisation dans $\mathbb{R}^n$ suivant :
  $$\text{(P. Primal)} \quad \begin{cases} \text{Minimiser} & f(x) \\ & g(x) \leq 0 \\ & x \in S \end{cases}$$
  où S est un sous-ensemble de $\mathbb{R}^n$, $f : S \to \mathbb{R}$ et $g : S \to \mathbb{R}^p$.
- Le Lagrangien du problème $L : S \times (\mathbb{R}^+)^p \to \mathbb{R}$:
  $$L(x,u) \quad = \quad f(x) + <u, g(x)> \quad \text{où} \quad u \in (\mathbb{R}^+)^p$$
  u est appelé le vecteur des multiplicateurs (de Lagrange)
- fonction duale associée au problème $h : (\mathbb{R}^+)^p \to \mathbb{R}$ telle que :
  $$h(u) = \inf_{x \in S} L(x,u) \quad \text{définie sur} \quad U = \{u \in (\mathbb{R}^+)^p \mid \inf_{x \in S} L(x,u) > -\infty\}.$$
- Théorème : La fonction h est concave sur tout sous-ensemble convexe de U.
- Théorème : $h(u) \leq f(x) \quad \forall u \in U \quad$ et $\quad \forall x \in S$ tel que $g(x) \leq 0$
- Si $\sup_h = +\infty$, le problème primal n'a pas de solution réalisable.
  Si $\inf_f = -\infty$, le problème dual n'a pas de solution réalisable.
- Le problème dual associé à (P) est :
  (P. Dual) $\quad$ Maximiser $h(u)$ pour $u \in U$

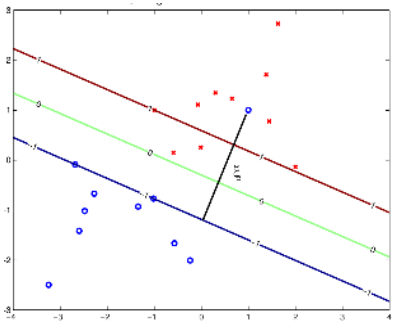
Formulation duale

- Formulation duale
  - Le dual est un problème quadratique de taille n (# observations).
  - $\exists$ algorithmes bien étudiés et très performants.
  - La formulation duale fait apparaître la matrice de Gram $X.X^t$, ce qui permet de gérer le cas non linéaire à travers des algorithmes à noyaux (qui seront étudiés dans la séance suivante).
- On introduit les multiplicateurs α_i de Lagrange. Le lagrangien est :
  $$L(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i[y_i(w^t.x_i + b) - 1]$$
  Le lagrangien doit être optimisé par rapport à w, b et les multiplicateurs α_i.
- En annulant les dérivées partielles du Lagrangien par rapport à w et b, on obtient les relations :
  $$\frac{\partial}{\partial b} L(w^*, b^*, \alpha^*) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i^*.y_i = 0$$
  $$\frac{\partial}{\partial w} L(w^*, b^*, \alpha^*) = 0 \quad \Rightarrow \quad w^* = \sum_{i=1}^{n} \alpha_i^*.y_i.x_i$$
- Par substitution de w par $\sum_{i=1}^{n} \alpha_i^*.y_i.x_i$ dans l'équation du Lagrangien, on obtient le problème dual :
  $$\begin{cases} max_\alpha \sum_{i=1}^{i=n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ t.q \\ \alpha_i \geq 0, i = 1, ..., n \quad \text{(admissibilité duale)} \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(stationarité)} \end{cases}$$

- La solution du problème dual donne les multiplicateurs de Lagrange optimaux α_i^*.
- A partir des α_i
  - on obtient w^* par les relations en haut.
  - Le paramètre b^* est obtenu à partir de la relation $|x_s^T w^* + b^*| = 1$ valable pour tous les vecteurs de support x_s.
- les vecteurs de support sont ceux pour lesquels $\alpha_i \geq 0$. En général leur nombre est beaucoup plus petit que le nombre total d'éléments dans la base d'apprentissage.
  $\Rightarrow$ Ajouter des échantillons qui ne sont pas des vecteurs supports à l'ensemble d'apprentissage n'a aucune influence sur la solution finale, c'est à dire seulement les vecteurs de support interviennent dans la fonction de décision (l'expression de la surface séparatrice entre les deux classes).
- Fonction de décision pour une nouvelle observation x :
  $$f^*(x) = \sum_{i=1}^{n} \alpha_i^* y_i x_i^T x + b^*$$
  L'hyperplan solution ne dépend que du produit scalaire entre le vecteur d'entrée et les vecteurs de support.
  $\Rightarrow$ Cette particularité permet l'utilisation de fonctions noyau pour aborder des problèmes non linéaires (traités dans la séance de cours suivante).

SVM : séparabilité ?

- Les SVM sont des séparateurs linéaires
  Que se passe-t-il si on dispose de données d'apprentissage non linéairement séparables ?



2 solutions :

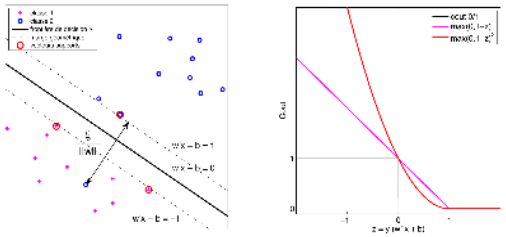
- Modifier le critère d'apprentissage de manière à autoriser des erreurs d'étiquetage (marge souple)
- Passer dans un espace de représentation où la séparabilité linéaire est possible : noyaux

SVM : Marge souple

- L'idée : modéliser les erreurs potentielles par des variables d'écart positives ξ_i associées aux observations $(x_i, y_i), i = 1, ...n$.
  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 coût de l'erreur) est nulle.
- Nous avons donc deux situations
  - Pas d'erreur : $y_i(w^T x_i + b) \geq 1 \Rightarrow \xi_i = 0$
  - Erreur : $y_i(w^T x_i + b) < 1 \Rightarrow \xi_i = 1 - y_i(w^T x_i + b) > 0$.

  On associe à cette définition une fonction coût appelée "coût charnière" :

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



Le pb d'optimisation dans le cas des données non-séparables est :

$$\begin{cases} min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^n \xi_i \\ \text{tel que} \quad \begin{cases} y_i(w.x_i + b) \geq 1 - \xi_i, i = 1, ..., n \\ \xi_i \geq 0, i = 1, ..., n \end{cases} \end{cases}$$

Si $\forall i, \xi_i = 0$, on retrouve le problème linéairement séparable traité plus tôt.

SVM : Marge souple

- L'idée : modéliser les erreurs potentielles par des variables d'écart positives ξ_i associées aux observations $(x_i, y_i), i = 1, ...n$.
  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 coût de l'erreur) est nulle.
- Nous avons donc deux situations
  - Pas d'erreur : $y_i(w^T x_i + b) \geq 1 \Rightarrow \xi_i = 0$
  - Erreur : $y_i(w^T x_i + b) < 1 \Rightarrow \xi_i = 1 - y_i(w^T x_i + b) > 0$.

  On associe à cette définition une fonction coût appelée "coût charnière" :

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

Le pb d'optimisation dans le cas des données non-séparables est :

$$\begin{cases} min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^n \xi_i \\ \text{tel que} \quad \begin{cases} y_i(w.x_i + b) \geq 1 - \xi_i, i = 1, ..., n \\ \xi_i \geq 0, i = 1, ..., n \end{cases} \end{cases}$$

Si $\forall i, \xi_i = 0$, on retrouve le problème linéairement séparable traité plus tôt.

SVM : Marge souple - Lagrangiens L_p et L_D

- On introduit les multiplicateurs α_i de Lagrange. Le lagrangien est :

$$L_P(w, b, \alpha) = \frac{1}{2}||w||^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i[y_i(w^t.x_i + b) - 1 + \xi_i] - \sum_{i=1}^n \mu_i \xi_i$$

Le lagrangien doit être optimisé par rapport à w, b et les multiplicateurs α_i et μ_i.

- En annulant les dérivées partielles du Lagrangien par rapport à w, b et ξ_i, on obtient les relations :

$$\begin{array}{lll} \frac{\partial L}{\partial b} L(w^*, b^*, \alpha^*) = 0 & \Rightarrow & \sum_{i=1}^n \alpha_i^*.y_i = 0 \\ \frac{\partial L}{\partial w} L(w^*, b^*, \alpha^*) = 0 & \Rightarrow & w^* = \sum_{i=1}^n \alpha_i^*.y_i.x_i \\ \frac{\partial L}{\partial \xi_i} L(w^*, b^*, \alpha^*) = 0 & \Rightarrow & \mu_i = C - \alpha_i \end{array}$$

- Par substitution de w par $\sum_{i=1}^n \alpha_i^*.y_i.x_i$ dans l'équation du Lagrangien, on obtient le problème dual :

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

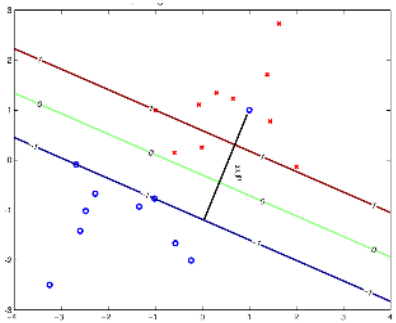
SVM : Marge souple - décision

- Observations :
  - C joue le rôle d'une constante de régularisation (la régularisation est d'autant plus forte que C est proche de 0).
    la régularisation : processus consistant à ajouter de l'information à un problème pour éviter le surapprentissage.
    - généralement, pénalité envers la complexité du modèle.
    - On peut relier cette méthode au principe du rasoir d'Occam.
    - D'un point de vue bayésien, l'utilisation de la régularisation revient à imposer une distribution a priori sur les paramètres du modèle.
  - La différence pour le problème dual 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 : séparabilité ?

- Les SVM sont des séparateurs linéaires
  Que se passe-t-il si on dispose de données d'apprentissage non linéairement séparables ?

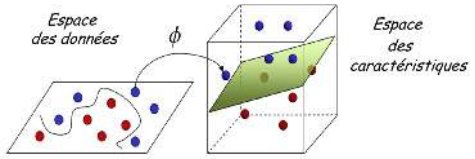


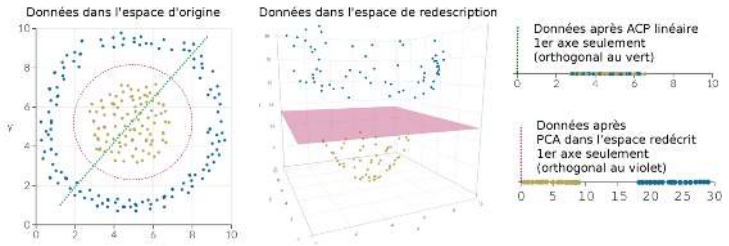
2 solutions :

- Modifier le critère d'apprentissage de manière à autoriser des erreurs d'étiquetage (marge souple)
- Passer dans un espace de représentation où la séparabilité linéaire est possible : noyaux

SVM : introduction au noyaux

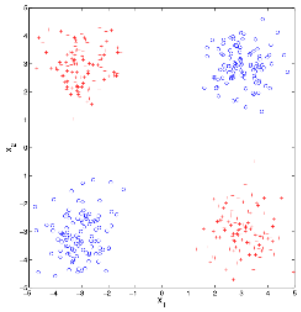
- Autre possibilité pour surmonter le problème des données non linéairement séparables dans l'espace d'entrée (input space) : Passer dans un nouvelle espace Φ (feature space) de grande dimension
- Un séparateur linéaire dans $\Phi(E)$ donne un séparateur non-linéaire dans E.

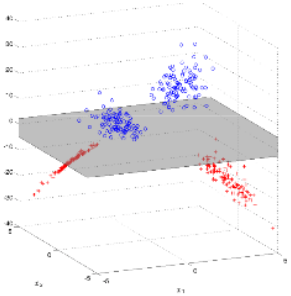




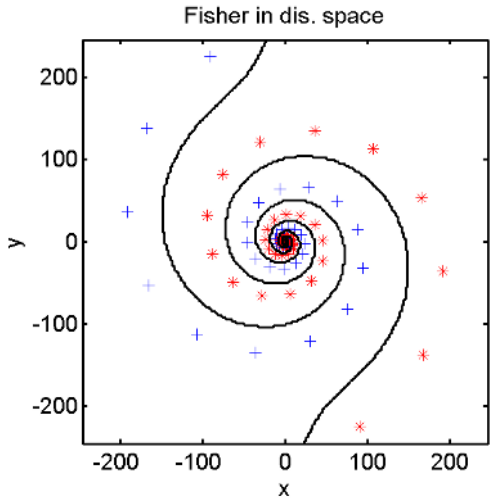
SVM : noyaux explicites

- input space : (x_1, x_2), en 2D
- feature space : $(x_1, x_2, x_1 x_2)$, en 3D





SVM : noyaux explicites



c'est un hyperplan, dans un certain espace...

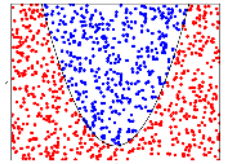
SVM : kernel trick

- Solution du SVM dans l'espace induit (feature space)

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i <\Phi(x_i), \Phi(x)> + b$$

- Au lieu de définir explicitement Φ, on préfère définir K :

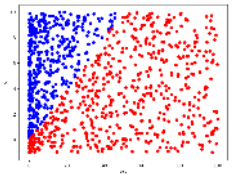
$$K(x, x') = <\Phi(x); \Phi(x')>$$



$\Rightarrow$ permet de faire des calculs dans l'espace de départ
Utile, surtout si $dim(\Phi) = \infty$

Exemple :

$$\begin{aligned} x &= (x_1, x_2) \\ \Phi(x) &= (x_1^2, \sqrt{2}x_1x_2, x_2^2) \\ \Phi(x)\Phi(x') &= x_1^2x'^2_1 + 2x_1x_2x'_1x'_2 + x_2^2x'^2_2 \\ &= (x_1x'_1 + x_2x'_2)^2 = (xx')^2 \end{aligned}$$



$\Rightarrow K(x, x') = (xx')^2$

Les noyaux : kernel trick

- On peut donc exprimer la solution SVM sans expliciter Φ

$$f(x) = \sum_{i \in \text{support}} \alpha_i y_i K(x_i, x) + b$$

- On peut choisir n'importe quelle fonction K
  Pourvu qu'on puisse prouver qu'il existe un espace dans lequel $K(x, x') = <\Phi(x), \Phi(x')>$
- Condition de Mercer :
  $k(x_i, x_j)$ terme général d'une matrice Semi Définie Positive (SDP)
  Valeur propres ≥ 0

Exemples de fonctions noyaux

- Linéaire : $K(x, x') = <x|x'>$
- Polynomiale : $K(x, x') = (<x|x'>)^d$ ou $= (< x|x' > +1)^d$
- Gaussienne (radial basis) : $K(x, x') = e^{-(||x-x'||2)/\sigma^2)}$

Qu'est-ce qu'un noyau ?

- Intuitivement, une fonction noyau est proche d'une mesure de similarité. *Si x et y sont 2 vecteurs, alors un noyau défini pour ces vecteurs doit avoir des valeurs élevées si $x \approx y$ et des valeurs faibles si x et y sont très différents.*
- Une fonction $s : X \times X \to \mathbb{R}$ est une mesure de similarité si :
  - $\forall x, y \in X, s(x, y) \geq 0$ (positivité)
  - $\forall x, y \in X, s(x, y) = s(y, x)$ (symétrie)
  - $\forall y \in X, y \neq x, s(x, y) < s(x, x)$ (uniformité)
  - $s(x, y) = s(x, x) \Leftrightarrow x = y$ (identité)

  A comparer avec le produit scalaire : deux vecteurs proches (directions presque parallèles) ont une valeur élevée pour le produit scalaire.

Théorème de Mercer.

- Théorème de Mercer. Soit X un compact dans $\mathbb{R}^d$ (compact = fermé et borné) et $K : X \times X \to \mathbb{R}$ une fonction symétrique. On suppose aussi que $\forall f \in L^2(X)$:

$$\sum_X K(x, y) f(x) f(y) dx dy \geq 0 \qquad \text{(condition de Mercer)}$$

  Alors il existe un espace de Hilbert H et $\Phi : X \to H$ tel que $\forall x, y \in X$:

$$K(x, y) = < \Phi(x), \Phi(y) > \qquad \text{(produit scalaire)}$$

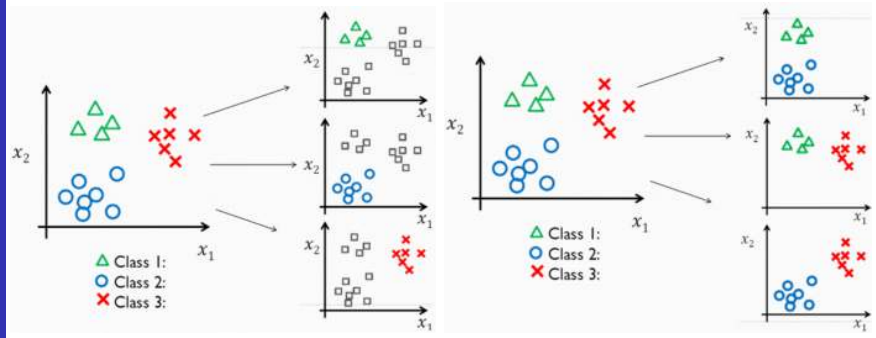
- La fonction K(x,y) s'appelle noyau positif défini.
- Une condition équivalente pour que la fonction $K : X \times X \to \mathbb{R}$ soit un noyau positif défini est la suivante :

$$\forall n \in N \text{ et } \{x_i\}_{i=1,\ldots,n} \subset X$$
$$\text{la matrice de Gramm} \qquad K = [K_{i,j}]_{i,j=1,\ldots,n} = [K(x_i, x_j)]_{i,j=1,\ldots,n}$$

  est définie positive, c'est à dire : $\forall c \in \mathbb{R}^n, c \neq 0$,on a $c^T K c > 0$
- Un noyau valide garantit donc l'existence de H et peut s'exprimer alors comme un produit scalaire dans H. Un noyau valide garantit aussi la convexité du problème d'optimisation quadratique sous contraintes d'inégalités rencontré pour les SVM.

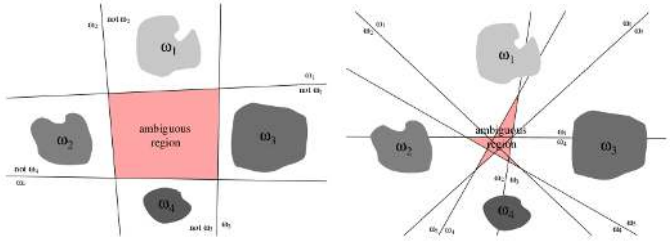
Classification multi-classe

- Solutions pour les problèmes multi-catégoriels :
  Convertir le problème en un ensemble de problèmes bi-classes
  - "one-versus-rest" / "one-against-all"
    Pour chaque classe C_i, on recherche une fonction discriminante qui sépare les exemples de la classes C_i de tous les autres exemples
  - "one-versus-one"
    $c(c-1)/2$ fonctions discriminantes sont utilisés, une par paire de classes



Classification multi-classe : ambiguité

- Convertir un problème multi-classes en un ensemble de problèmes biclasses peut mener à des régions dans lesquelles la classification est indéfinies...



- OvR : on choisit la classe prédite avec le meilleur score
- OvO : on choisit la classe qui remporte le plus de duels.

♡ Avantage d'OvO : taille des ensembles d'apprentissage faible, très utile pour SVM car SVM ne passe pas bien à l'échelle