Introduction aux Séparateurs à Vaste Marges

(Support Vector Machines (SVM))

Youssef SALMAN
youssef.salman@emse.fr

Ecole des Mines de Saint-Étienne
Majeure: Sciences de données

17 novembre 2025

Table of contents

Contexte

Problème d'apprentissage

On s'intéresse à un phénomène f (déterministe ou stochastique) qui,

- à partir d'un certain jeu d'entrées x (variables),
- produit une sortie $y = f(x)$ (label, étiquette).

GOAL

Le but est à partir de la seule observation d'un certain nombre de couples entrée-sortie $\{(x_i, y_i)\} n_i = 1$

Apprentissage automatique : Définition et formalisation

- L'apprentissage automatique vise à :
  - construire un modèle général
  - à partir de données particulières.
- dans le but de :
  - prédire un comportement face à une nouvelle donné
  - approximier une fonction ou une densité de probabilité
- **Données** : $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ appelé **ensemble d'apprentissage**.
  - Chaque x_i est un vecteur dans l'espace $\mathcal{X}$ (par exemple, $\mathcal{X} = \mathbb{R}^3$).
  - Dans ce cours, on s'intéresse à la classification et donc au cas où $\mathcal{Y} = \{-1, 1\}$.
- **But** : construire une fonction appelée classifieur $f : \mathcal{X} \to \{-1, 1\}$ qui minimise la probabilité de classification :

$$\mathbb{P}\big(f(x) \neq y\big).$$

Support Vector Machines

Qu'est-ce qu'un SVM ?

Definition

Les **Séparateurs à Vaste Marges** également appelés **Machines à Vecteurs de Support** (*Support Vector Machine* ou **SVM**) sont des algorithmes d'apprentissage supervisé destinés à résoudre des problèmes de classification ou de régression.

- Ils ont été introduits par **Vladimir Vapnik (1992)**.
  - Vapnik V. N. *The Nature of Statistical Learning Theory.*
  - Boser B.E, et al. *A Training Algorithm for Optimal Margin Classifiers*, Fifth Annual Workshop on Computational Learning Theory.
- Dans ce cours, nous verrons le problème de **classification**

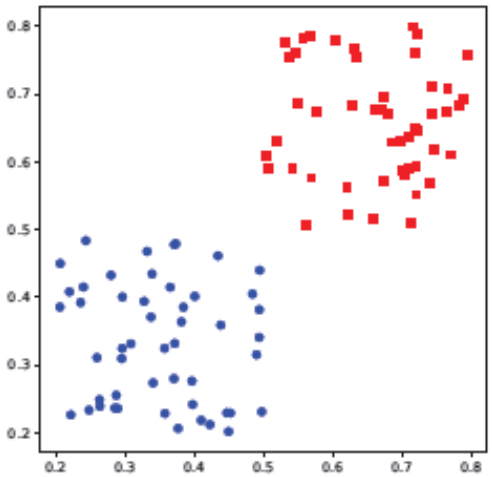
Illustration SVM dans le cas d'un problème de classification

- **Rappel du problème d'apprentissage**
  - On s'intéresse à un phénomène f (déterministe ou stochastique) qui à partir d'un certain jeu d'entrée x (variables) produit une sortie y (label, étiquette) : $y = f(x)$
  - L'objectif est de trouver la fonction f (inconnue) à partir d'un échantillon d'apprentissage $S = \{(x_i, y_i)\}_{i=1}^{n} \in \mathcal{X} \times \mathcal{Y}$

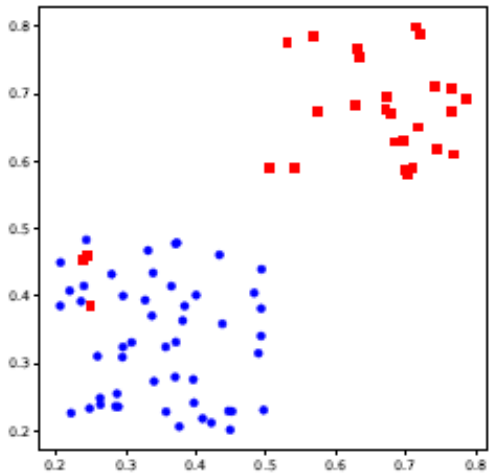
- **Les données : échantillon d'apprentissage**
  - Chaque x_i est un vecteur dans l'espace $\mathcal{X}$
  - Dans le cas de la classification on s'intéresse au cas où $\mathcal{Y} = \{-1, 1\}$

Données linéairement séparables

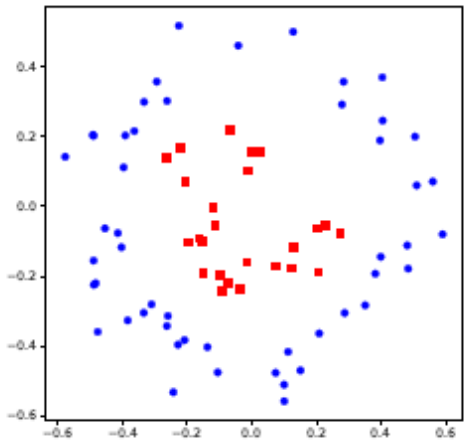
- Exemple d'échantillon de donnée en 2 dimensions : $\mathcal{X} = \mathbb{R}^2$
- Quel échantillon est linéairement séparable ?



Linéairement séparable



Presque linéairement séparable



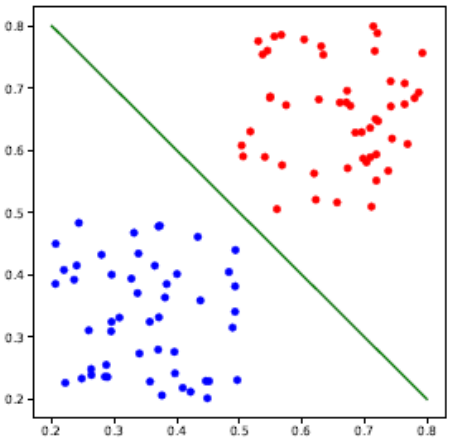
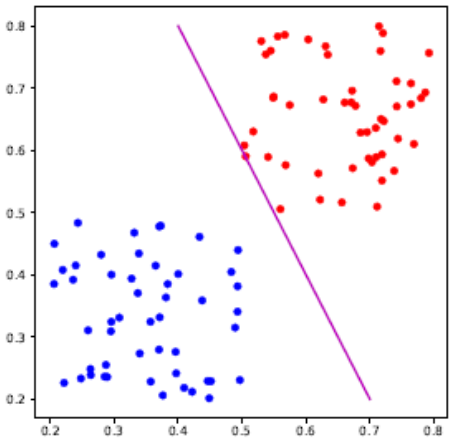
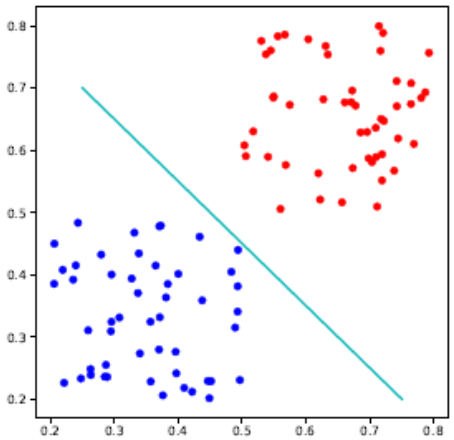
Non linéairement séparable

GOAL

Comment trouver le "meilleur" séparateur linéaire ?

Comment trouver le meilleur séparateur linéaire ?

Laquelle parmi ces droites sépare bien les deux classes ? Pourquoi ?

Notion d'hyperplan

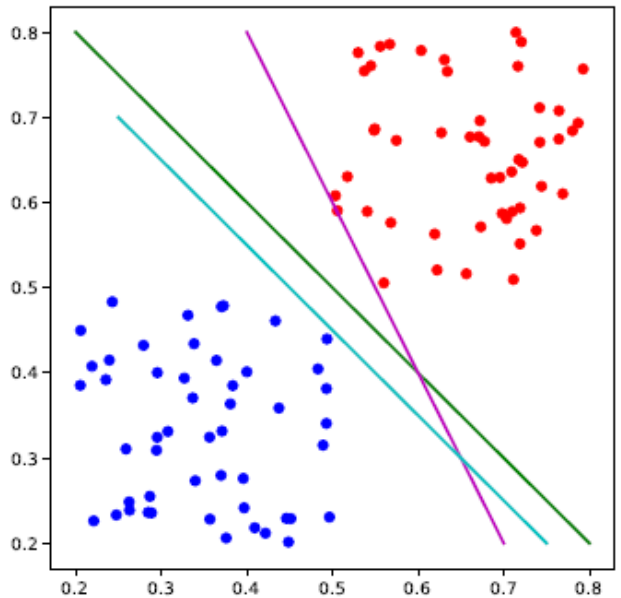
Un hyperplan séparateur H sépare l'espace des données d'entrées $\mathcal{X}$ en deux demi-espaces correspondant aux deux classes prévues pour y.

- En dimension 2, hyperplan = droite
- En dimension d, $\mathbf{x} = (x[1] \ \cdots \ x[d])^T$, un hyperplan est défini par :

$$\beta_1 x[1] + \beta_2 x[2] + \cdots + \beta_d x[d] = \langle \boldsymbol{\beta}, \mathbf{x} \rangle = -\beta_0$$

- $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_n)^T$ est le vecteur des paramètres.
- On le note :

$$h(x) = \langle \boldsymbol{\beta}, \mathbf{x} \rangle + \beta_0$$



Hyperplan séparateur

Définition

Un hyperplan séparateur H sépare l'espace des données d'entrées $\mathcal{X}$ en deux demi-espaces correspondant aux deux classes prévues pour y. Il est défini par l'équation $h(\mathbf{x}) = 0$:

$$H = \{\mathbf{x} \mid \langle \boldsymbol{\beta}, \mathbf{x} \rangle + \beta_0 = 0\}$$

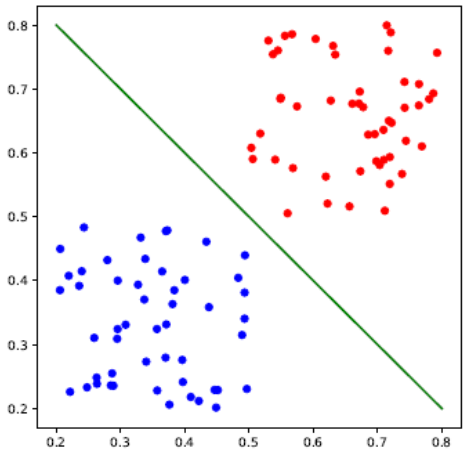


Figure – Hyperplan en 2D

SVM

Remarque

- *La droite verte est aussi appelée **frontière de décision (FD)** ou decision boundary dans le cas d'un SVM linéaire (comme ici).*
- *La FD permet de classifier à droite comme des rouges donc +1 et à gauche comme des bleues donc -1.*
- *Elle doit être suffisamment loin des deux classes. Pour cela on maximise ce qu'on appelle la marge.*

Classifieur et fonction de décision

1. Qu'est-ce qu'un classifieur ?

- Un classifieur est une fonction

$$f : \mathcal{X} \to \{-1, 1\}$$

qui associe à chaque observation x une étiquette :

$$f(x) = \begin{cases} +1 & \text{si } x \text{ appartient à la classe rouge} \\ -1 & \text{si } x \text{ appartient à la classe bleue} \end{cases}$$

2. Minimiser la probabilité d'erreur

- L'objectif est de trouver le classifieur f qui fait le moins d'erreurs possibles, c'est-à-dire qui minimise :

$$\mathbb{P}(f(x) \neq y)$$

où y est la vraie classe de l'observation x.

Classifieur linéaire

Definition

Un **classifieur linéaire** est une fonction de la forme :

$$f(x) = \text{sgn}(h(x)) = \begin{cases} +1 & \text{si } h(x) > 0, \\ 0 & \text{si } h(x) = 0 \Leftrightarrow x \in H, \\ -1 & \text{si } h(x) < 0, \end{cases}$$

où la **fonction de décision** est donnée par :

$$h(x) = \langle \boldsymbol{\beta}, \mathbf{x} \rangle + \beta_0$$

avec $\boldsymbol{\beta} \in \mathbb{R}^d$ et $\beta_0 \in \mathbb{R}$.

Interprétation géométrique

Remarque (Interprétation géométrique)

- *Le vecteur β détermine l'**orientation** de la frontière de décision.*
- *Le biais β_0 contrôle son **positionnement** dans l'espace.*
- *Le plan $H = \{x \mid h(x) = 0\}$ représente la **frontière de décision** : il sépare les points pour lesquels $f(x) = +1$ (classe rouge) et $f(x) = -1$ (classe bleue).*

Interprétation géométrique

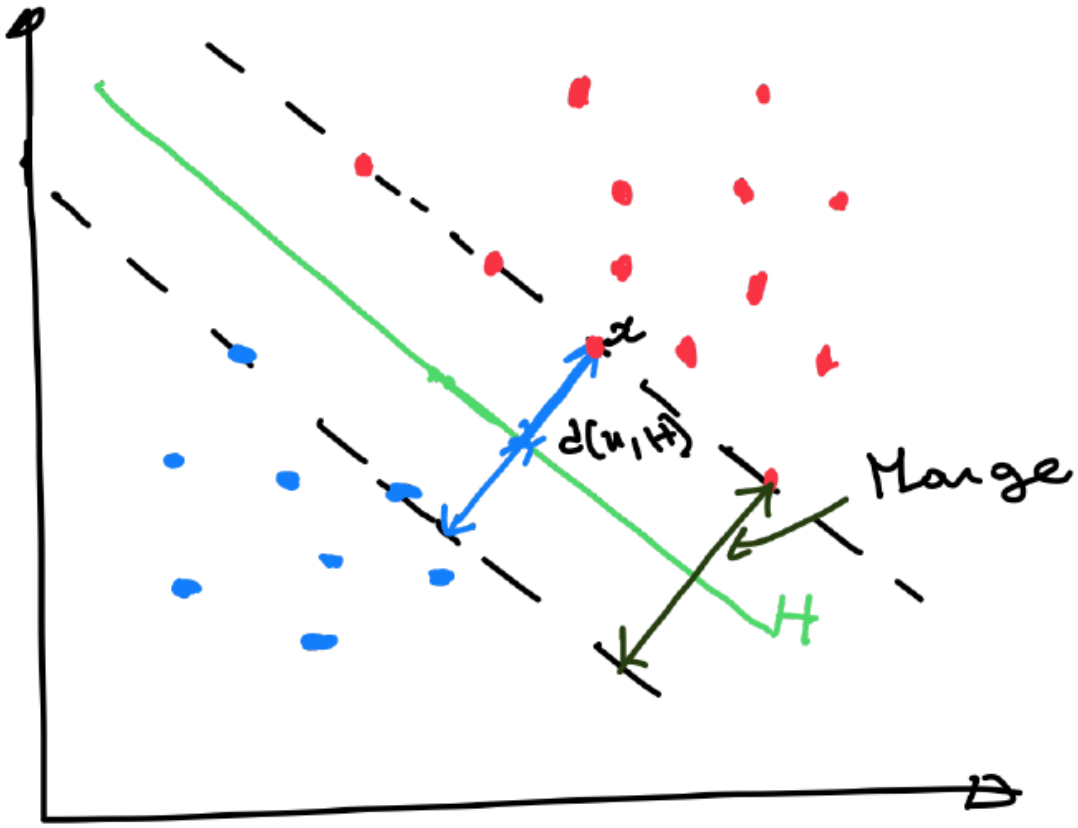
Remarque (Interprétation géométrique)

- *Le vecteur β détermine l'**orientation** de la frontière de décision.*
- *Le biais β_0 contrôle son **positionnement** dans l'espace.*
- *Le plan $H = \{x \mid h(x) = 0\}$ représente la **frontière de décision** : il sépare les points pour lesquels $f(x) = +1$ (classe rouge) et $f(x) = -1$ (classe bleue).*

Remarque

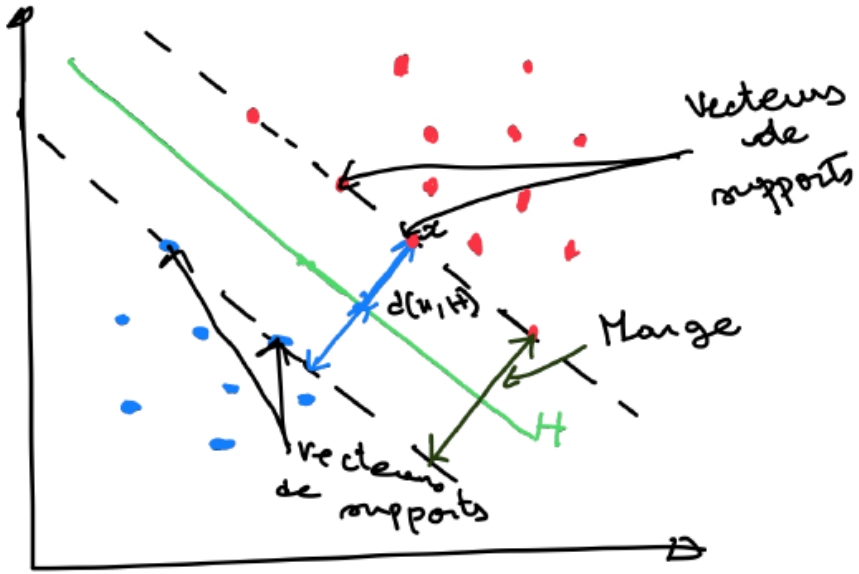
- *Si toutes les observations vérifient $f(x) = +1$ dans la zone rouge et $f(x) = -1$ dans la zone bleue, alors il n'y a **aucune erreur de classification**.*
- *Dans ce cas, il existe un **classifieur linéaire** qui sépare parfaitement les deux classes : on dit qu'elles sont **linéairement séparables**.*

La marge



$$\text{Marge} = 2d(\boldsymbol{x}, H)$$

Vecteurs de support et objectif



- Certaines observations de nos données d'entraînement ont un nom spécial : on les appelle les vecteurs de support.
- Qu'est-ce qui les rend si spéciales ?

Pourquoi ces points sont-ils spéciaux ?

Remarque

Les vecteurs de support sont les points d'entraînement les plus proches de la frontière de décision.

- Ce sont eux qui **déterminent entièrement l'hyperplan optimal**.
- Déplacer un point non support ne change pas la solution du SVM.
- En revanche, **déplacer un vecteur de support modifie la marge et donc l'hyperplan**.

Objectif : trouver l'hyperplan **optimal** au sens de la marge (parmi une infinité).

Cas de SVM Linéaire

Deux cas de SVM linéaires :

- SVM linéaire à marge dure (hard margin linear SVM)
- SVM linéaire à marge souple (soft margin linear SVM)

SVM à Marge Dure

SVM à marge dure

Rappels

Soient $\mathbf{u}$ et $\mathbf{v}$ deux vecteurs de $\mathbb{R}^d$:

- Produit scalaire :

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{j=1}^{d} u_j v_j \quad \text{et on notera } \langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v}.$$

- Norme euclidienne :

$$\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}.$$

- Vecteur normal à un hyperplan H :

$$\tilde{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}.$$

- Orthogonalité : $\mathbf{u}$ et $\mathbf{v}$ sont orthogonaux si $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.
- Distance d'un point $\mathbf{x} \in \mathbb{R}^d$ à l'hyperplan H :

$$d(\mathbf{x}, H) = |\langle \tilde{\boldsymbol{\beta}}, \mathbf{x} - \mathbf{x}_0 \rangle| \quad \text{avec } \mathbf{x}_0 \in H.$$

Construction de la règle de décision

- Soit **x** (une nouvelle variable ou donnée) un point inconnu de l'espace

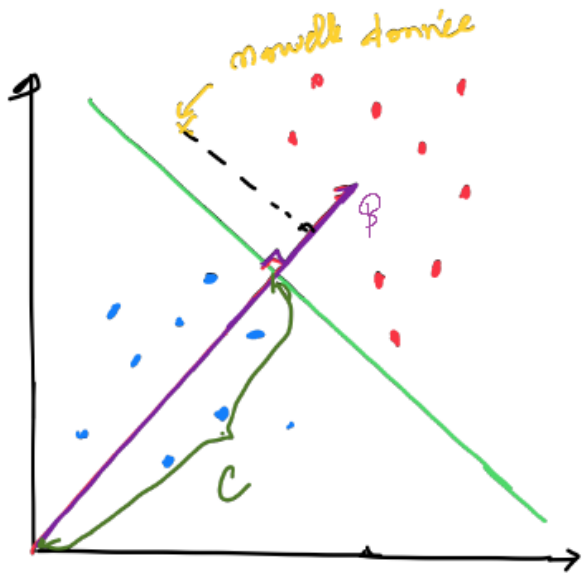
$$\mathcal{X} = \mathbb{R}^d.$$

- Question : **x** est-il à **gauche** ou à **droite** de la droite (ou de l'hyperplan) de décision ?

Construction de la règle de décision

Pour classer une nouvelle donnée **x** :

- On trace un vecteur β perpendiculaire à la droite de décision.
- On appelle c la longueur de l'origine du repère au point d'intersection avec la droite de décision.
- On projette **x** sur β : il s'agit du produit scalaire entre **x** et β.
- On calcule $\langle \beta, \mathbf{x} \rangle = \beta \cdot \mathbf{x}$ et on le compare à c.



Règle de décision

- si $\beta \cdot \mathbf{x} = c$ alors $\mathbf{x}$ est sur la droite de décision
- si $\beta \cdot \mathbf{x} < c$ alors $\mathbf{x}$ est un point bleu
- si $\beta \cdot \mathbf{x} > c$ alors $\mathbf{x}$ est un point rouge

Plus généralement, on calcule $\beta \cdot \mathbf{x} + b$ avec $b = -c$ et β inconnues.

Normalisation et hyperplan canonique

En général, on normalise les paramètres β et b de sorte que la **fonction de décision** évaluée sur un vecteur de support soit égale à $+1$ ou -1. On obtient ainsi un hyperplan canonique.

Hyperplan canonique

Un hyperplan $\beta \cdot \mathbf{x} + b = 0$ est dit **canonique** selon $\mathcal{S}$ si le plus proche point de $\mathcal{S}$ est à la distance unitaire de cet hyperplan :

$$\min_{\mathbf{x}_i \in \mathcal{S}} |\beta \cdot \mathbf{x}_i + b| = 1$$

Conditions d'une classification parfaite

Pour un **hyperplan canonique**, les conditions d'une classification parfaite (sur l'échantillon d'apprentissage) sont de la forme :

- $h(x_i) = \beta \cdot \mathbf{x}_i + b \geq +1$ si l'on se trouve à droite (a)
- $h(x_i) = \beta \cdot \mathbf{x}_i + b \leq -1$ si l'on se trouve à gauche (b)

$$\text{comme} \quad y_i = \begin{cases} +1 & \text{si rouge,} \\ -1 & \text{si bleu,} \end{cases} \quad \Rightarrow \quad \text{alors (a) et (b) deviennent :}$$

$$y_i(\beta \cdot \mathbf{x}_i + b) \geq 1, \quad \text{i.e.} \quad y_i(\beta \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad (1)$$

et

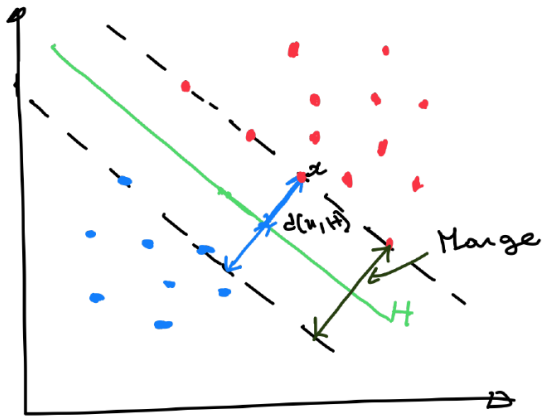
$$y_i(\beta \cdot \mathbf{x}_i + b) - 1 = 0 \quad (2)$$

pour les $\mathbf{x}_i$ qui sont des **vecteurs de support**.

La marge et sa largeur I

Definition

Marge=double de la distance du point le plus proche de l'hyperplan



$$\text{Marge} = 2d(\boldsymbol{x}, H)$$

Soient deux vecteurs $\mathbf{x}_r$ et $\mathbf{x}_b$ **situés de part et d'autre de la marge**. Soit β le vecteur unitaire orthogonal à l'hyperplan.

- Quel est la **largeur de la marge** ?

La marge et sa largeur II

- largeur de la marge $= \mathbf{x}_r - \mathbf{x}_b$
- Comme $\tilde{\beta}$ est unitaire, on peut considérer :

$$\text{largeur} = \frac{\beta}{\|\beta\|} \cdot (\mathbf{x}_r - \mathbf{x}_b) = \frac{\beta}{\|\beta\|} \cdot \mathbf{x}_r - \frac{\beta}{\|\beta\|} \cdot \mathbf{x}_b$$

- Pour $\mathbf{x}_r$, on a $y_i = +1$, donc :

$$y_i(\beta \cdot \mathbf{x}_i + b) = 1 \;\Rightarrow\; \beta \cdot \mathbf{x}_i = 1 - b$$

- Pour $\mathbf{x}_b$, on a $y_i = -1$, donc :

$$y_i(\beta \cdot \mathbf{x}_i + b) = 1 \;\Rightarrow\; \beta \cdot \mathbf{x}_i = -(1 + b)$$

- La largeur de la marge devient :

$$\text{largeur} = \frac{1 - b + 1 + b}{\|\beta\|} = \frac{2}{\|\beta\|}$$

La marge et sa largeur III

Ainsi, la marge d'un hyperplan canonique séparateur est égale à :

$$\boxed{\frac{2}{\|\beta\|}}$$

Rappel

Definition (Lagrangien)

Le Lagrangien est une fonction qui permet d'intégrer les contraintes d'un problème d'optimisation dans la fonction objectif, au moyen de multiplicateurs appelés multiplicateurs de Lagrange.
Pour un problème :

$$\min_{\beta} \; f(\beta) \quad \text{s.c.} \quad g_i(\beta) \geq 0,$$

le Lagrangien est :

$$L(\beta, \alpha) = f(\beta) - \sum_i \alpha_i g_i(\beta), \quad \text{avec } \alpha_i \geq 0.$$

Formulation du problème primal d'optimisation

On veut **maximiser la marge**, donc on veut maximiser

$$\frac{2}{\|\boldsymbol{\beta}\|}$$

ce qui est équivalent à **minimiser** $\|\boldsymbol{\beta}\|$:

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

en respectant les **contraintes linéaires** suivantes :

$$y_i(\boldsymbol{\beta} \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad \forall i = 1, \ldots, n$$

SVM : un problème d'optimisation quadratique I

Formulation du problème primal d'optimisation :

$$\min_{\beta,\, b} \frac{1}{2}\|\beta\|^2 \quad \text{s.c.} \quad y_i(\beta \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad \forall i = 1, \ldots, n$$

- **Problème d'optimisation quadratique convexe** avec des contraintes linéaires.
- Pas de **solution analytique**.
- Plusieurs méthodes de résolution : gradient conjugué, méthode du point intérieur, etc.
- Non opérationnel si $d \gg n$ grand.
- Valable seulement si les données sont **linéairement séparables**.

SVM : un problème d'optimisation quadratique II

- **Solution dans $\mathbb{R}^n$ (convexe) : forme duale** Intégrer les contraintes linéaires dans la fonctionnelle. Si $\boldsymbol{\beta} \in \mathbb{R}^d$ et $\boldsymbol{\alpha} \in \mathbb{R}^n_+$,

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) := \phi(\boldsymbol{\beta}) + \boldsymbol{\alpha}^\top \psi(\boldsymbol{\beta}),$$

  où la fonction $\phi(\cdot)$ représente la **fonction de coût** et $\psi(\cdot)$ représente les n **contraintes**.
- Méthode éprouvée : **conditions de Karush–Kuhn–Tucker (KKT)**.
- Le problème d'optimisation étant **convexe**, il est **équivalent de résoudre le problème primal ou le problème dual**.

Les multiplicateurs de Lagrange I

Lagrangien

Le Lagrangien associé au programme du SVM s'écrit :

$$\mathcal{L}(\beta, b, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\beta \cdot \mathbf{x}_i + b) - 1 \right]$$

Les α_i sont les multiplicateurs de Lagrange, **un par contrainte**. Ils vérifient, pour tout $i = 1, \ldots, n$:

$$\alpha_i \geq 0$$

Remarque

- *Le Lagrangien doit être **minimisé** par rapport à β et b, et **maximisé** par rapport à α.*
- *Si $[y_i(\beta \cdot \mathbf{x}_i + b) - 1] = 0$, alors $\alpha_i > 0$ et $\mathbf{x}_i$ est un **vecteur de support**.*
- *Si $[y_i(\beta \cdot \mathbf{x}_i + b) - 1] > 0$, alors $\alpha_i = 0$ et le point $\mathbf{x}_i$ situé au-delà de la marge est bien classé. En pratique, beaucoup de α_i sont nuls.*

Formulation duale du programme SVM

On rappelle le problème primal :

$$\min_{\beta,\, b} \frac{1}{2}\|\beta\|^2 \quad \text{s.c.} \quad y_i(\beta \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad \forall i = 1, \ldots, n$$

Les solutions β et b peuvent être obtenues en résolvant le problème dual suivant :

$$\min_{\beta,\, b} \max_{\alpha \geq 0} \mathcal{L}(\beta, b, \alpha)$$

Permutons le min et le max :
On peut montrer que $\mathcal{L}$ est convexe selon β et b, et concave selon α, ce qui rend possible :

$$\min_{\beta,\, b} \max_{\alpha \geq 0} \mathcal{L}(\beta, b, \alpha) = \max_{\alpha \geq 0} \min_{\beta,\, b} \mathcal{L}(\beta, b, \alpha)$$

Conditions de Karush-Kuhn-Tucker (KKT)

Annuler le gradient

$$\min_{\beta,b} \mathcal{L}(\beta, b, \alpha) = \min_{\beta,b} \left(\frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\beta \cdot \mathbf{x}_i + b) - 1 \right] \right)$$

C'est un **problème d'optimisation strictement convexe**. Il suffit de trouver les bons β et b qui annulent le gradient de la fonctionnelle pour obtenir la solution.

Definition (Conditions KKT)

Les conditions de **Karush–Kuhn–Tucker (KKT)** sont définies par :

- $\dfrac{\partial \mathcal{L}(\beta, b, \alpha)}{\partial \beta} = 0 \Rightarrow \beta = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$
- $\dfrac{\partial \mathcal{L}(\beta, b, \alpha)}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$
- $\alpha_i [\, y_i(\beta \cdot \mathbf{x}_i + b) - 1 \,] = 0$
- $\alpha_i \geq 0, \quad \forall i = 1, \ldots, n$

Un programme quadratique dual

En injectant la valeur optimale de β et la contrainte comme solution du *min*, on obtient le problème suivant :

Problème dual

Le problème du SVM s'écrit sous forme duale comme :

$$\boldsymbol{\alpha}^* = \max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \, \mathbf{x}_i \cdot \mathbf{x}_j \right)$$

$$\text{s.c.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{et} \quad \boldsymbol{\alpha} \geq 0$$

Remarques

- Autant de variables d'optimisation que d'exemples d'apprentissage !
- seuls des produits scalaires entre des exemples d'apprentissage apparaissent !

Solution optimale du SVM I

- Une fois que l'on a les multiplicateurs optimaux $\boldsymbol{\alpha}^*$, on en déduit le vecteur des coefficients $\boldsymbol{\beta}^*$ comme :

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i = \sum_{i \in \text{Supp}} \alpha_i^* y_i \mathbf{x}_i$$

où Supp désigne l'ensemble des vecteurs de support, puisque $\alpha_i^* = 0$ pour $i \notin$ Supp.

- Pour déterminer la constante b^*, on considère n'importe quel vecteur de support $i \in$ Supp tel que :

$$y_i(\boldsymbol{\beta}^* \cdot \mathbf{x}_i + b^*) = 1$$

On obtient alors :

$$b^* = \frac{1}{y_i} - \boldsymbol{\beta}^* \cdot \mathbf{x}_i = y_i - \boldsymbol{\beta}^* \cdot \mathbf{x}_i \quad (\text{car } y_i \in \{-1, 1\}, \forall i \in \text{Supp})$$

Solution optimale du SVM II

- Pour tout point $\mathbf{x} \in \mathcal{X}$, la fonction de décision optimale $h(x)$ est :

$$h(x) = \beta^* \cdot \mathbf{x} + b^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} + b^*$$

Remarque

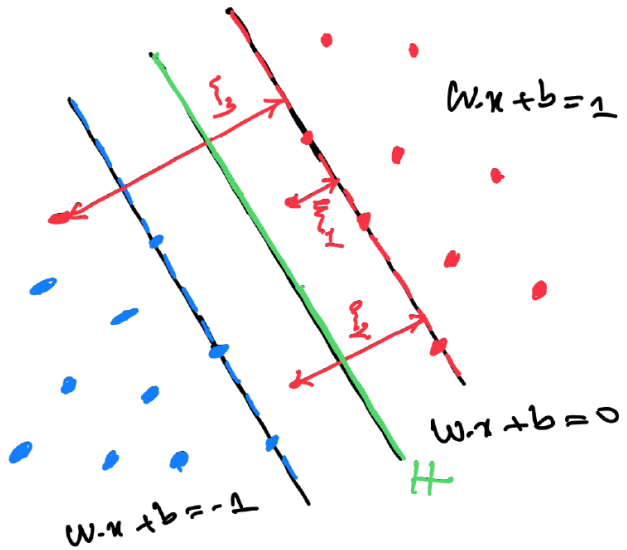
solution à base de vecteurs de support

- *Seuls les vecteurs de support (pondérés par les $\alpha_i \neq 0$) déterminent le classement de toutes les observations d'apprentissage et de test.*
- *C'est une solution **parcimonieuse**.*
- *Bonne résistance au bruit (marges douces...).*

SVM à Marge douce

SVM à marge douce

- Souvent, il arrive que même si le problème est linéaire, les données soient affectées par un bruit (par exemple un bruit de capteur), et les deux classes se retrouvent mélangées autour de l'hyperplan de séparation.
- Pour gérer ce type de problème, on utilise une technique dite de **marge souple**, qui tolère les mauvais classements :
  - On introduit des variables de relâchement de contraintes, notées ξ_i.
  - On pénalise ces variables de relâchement dans la fonction objectif.



Variables de relâchement (*slack variables*) I I

Variables de relâchement

Dans le cas d'un échantillon presque linéairement séparable, on introduit n variables de relaxation des contraintes de classification

$$y_i(\beta \cdot \mathbf{x}_i + b) \geq 1$$

Ces variables, notées

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_n)^\top,$$

sont appelées *variables de relâchement* (*slack variables*) et vérifient :

$$y_i(\beta \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$$

Cortes et al., *Support-Vector Networks*, Machine Learning.

$$y_i(\beta \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$$

avec $\xi_i \geq 0$. Trois configurations peuvent apparaître suivant les valeurs de ξ_i :

Variables de relâchement (*slack variables*) I II

1. Si $\xi_i = 0$, alors $y_i(\beta \cdot \mathbf{x}_i + b) \geq 1$: l'observation $(\mathbf{x}_i, y_i)$ est **bien classée**.
2. Si $\xi_i \geq 1$, alors l'observation $(\mathbf{x}_i, y_i)$ est située du mauvais côté de la frontière. On a :

$$y_i(\beta \cdot \mathbf{x}_i + b) < 1 \Rightarrow \xi_i = 1 - y_i(\beta \cdot \mathbf{x}_i + b) > 0$$

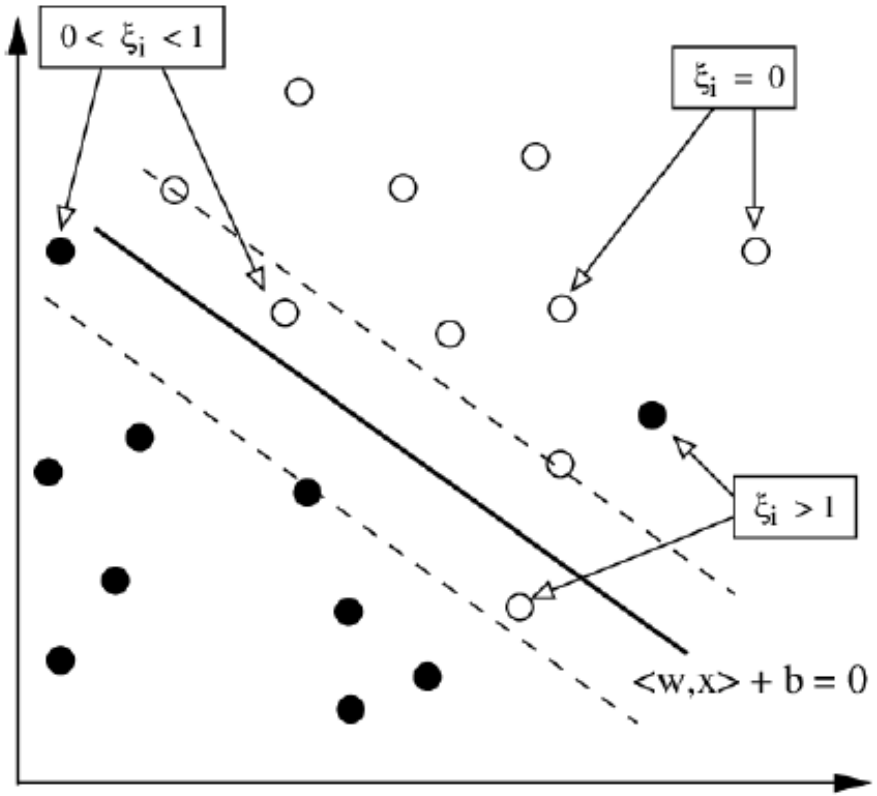
3. On associe aux deux points ci-dessus une fonction de coût appelée **coût charnière** (ou *Hinge cost* en anglais) :

$$\xi_i = \max(0,\ 1 - y_i(\beta \cdot \mathbf{x}_i + b))$$

4. Si $0 < \xi_i < 1$: l'observation $(\mathbf{x}_i, y_i)$ est nécessairement du bon côté de la frontière (bien classée), mais peut être située à une distance de l'hyperplan séparateur inférieure à la moitié de la marge :

$$d(\mathbf{x}_i, H) < \frac{1}{\|\beta\|}$$

Variables de relâchement (*slack variables*) I III



Formulation du problème primal I

L'objectif est de maximiser la marge tout en minimisant la somme des ξ_i. Le problème primal du SVM dans le cas des données non-séparables est donc :

$$\min_{\boldsymbol{\beta}, b, \boldsymbol{\xi} \geq 0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{s.c.} \quad \begin{cases} y_i(\boldsymbol{\beta} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, & i = 1, \ldots, n \\ \xi_i \geq 0, & i = 1, \ldots, n \end{cases}$$

où C désigne un *coefficient de pénalisation*.

Remarques

Remarque

$$\min_{\beta,\, b,\, \xi \geq 0} \underbrace{\frac{1}{2}\|\beta\|^2}_{\text{marge}} + \underbrace{C}_{\text{pénalité}} \underbrace{\sum_{i=1}^{n} \xi_i}_{\text{erreurs de classification}} \quad \text{s.c.} \quad \begin{cases} y_i(\beta \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \quad i = 1, \ldots, n \end{cases}$$

- *C joue le rôle d'une constante de régularisation. Il contrôle l'arbitrage entre la largeur de la marge et le taux d'erreur.*
- ***Si** C **est petit**, les erreurs de classification sont moins pénalisées et l'accent est mis sur **la maximisation de la marge**.*
  - *Il y a un risque de sous-apprentissage (mauvais taux de classification sur l'échantillon d'apprentissage).*
- ***Si** C **est grand**, l'accent est mis sur **l'absence de mauvaise classification au prix d'une marge plus faible**.*
  - *Il y a un risque de sur-apprentissage.*

Lagrangien

Dans le cas non-linéairement séparable, le Lagrangien associé au problème du SVM s'écrit :

$$\mathcal{L}(\beta, b, \xi, \alpha, \lambda) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[\, y_i(\beta \cdot \mathbf{x}_i + b) - 1 + \xi_i\,] - \sum_{i=1}^{n}\lambda_i\xi_i$$

Comme précédemment (*SVM à marge dure*), on a :

- $\dfrac{\partial \mathcal{L}(\beta, b, \xi, \alpha, \lambda)}{\partial \beta} = 0 \;\Rightarrow\; \beta = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i$ (identique au cas marge dure)
- $\dfrac{\partial \mathcal{L}(\beta, \xi, \alpha, \lambda)}{\partial b} = 0 \;\Rightarrow\; \sum_{i=1}^{n}\alpha_i y_i = 0$
- $\dfrac{\partial \mathcal{L}(\beta, b, \xi, \alpha, \lambda)}{\partial \xi_i} = 0 \;\Rightarrow\; C - \alpha_i - \lambda_i = 0 \;\Rightarrow\; C = \alpha_i + \lambda_i$

Formulation duale

En injectant les résultats obtenus lors de la minimisation par rapport à β, b et ξ, le problème dual s'écrit :

$$\alpha^* = \max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \, \mathbf{x}_i \cdot \mathbf{x}_j \right)$$

$$\text{s.c.} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0, \\ 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \end{cases}$$

Remarque : La différence pour le problème dual entre le cas séparable et le cas non séparable est que **les valeurs des α_i sont majorées par C**.

Remarques

Remarque

Trois cas possibles suivant les valeurs de α_i :

1. *Si $\alpha_i = 0$, alors $y_i(\beta \cdot \mathbf{x}_i + b) \geq 1$, $\lambda_i = C$, $\xi_i = 0$: l'observation $(\mathbf{x}_i, y_i)$ est **bien classée**.*
2. *Si $0 < \alpha_i < C$, alors $y_i(\beta \cdot \mathbf{x}_i + b) = 1$, $\lambda_i = C - \alpha_i > 0$, $\xi_i = 0$: l'observation $(\mathbf{x}_i, y_i)$ est un **vecteur de support**.*
3. *Si $\alpha_i = C$, alors $y_i(\beta \cdot \mathbf{x}_i + b) = 1 - \xi_i$, et $\xi_i > 0$ car $\lambda_i = C - \alpha_i > 0$: l'observation peut être située **dans la marge (bien classée)** ou **du mauvais côté (mal classée)**.*

Solution optimale du SVM

Une fois obtenus les multiplicateurs optimaux α_i^*, on en déduit le vecteur des coefficients β^* et b^* comme :

$$\beta^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i = \sum_{i \in \text{Supp}} \alpha_i^* y_i \mathbf{x}_i$$

et

$$b^* = y_i - \beta^* \cdot \mathbf{x}_i$$

où *Supp* désigne l'ensemble des vecteurs de support tels que $0 < \alpha_i < C$.

Paramètre de pénalisation

- En pratique, les performances du SVM sont très sensibles au choix du paramètre de pénalisation C.
- Comment sélectionner un hyperparamètre C optimal ?
  - Validation croisée
  - K-fold cross validation

Astuce de noyau

Pourquoi introduire l'astuce du noyau ?

Limite du SVM linéaire (marge dure ou douce)
Le SVM linéaire ne peut séparer que des données **linéairement séparables** dans l'espace d'entrée X.

Problème

Dans de nombreux cas réels :

- la frontière entre les classes est **non linéaire**,
- même une marge souple (ξ_i) ne suffit pas,
- aucune droite (hyperplan) ne sépare correctement les données.

Idée

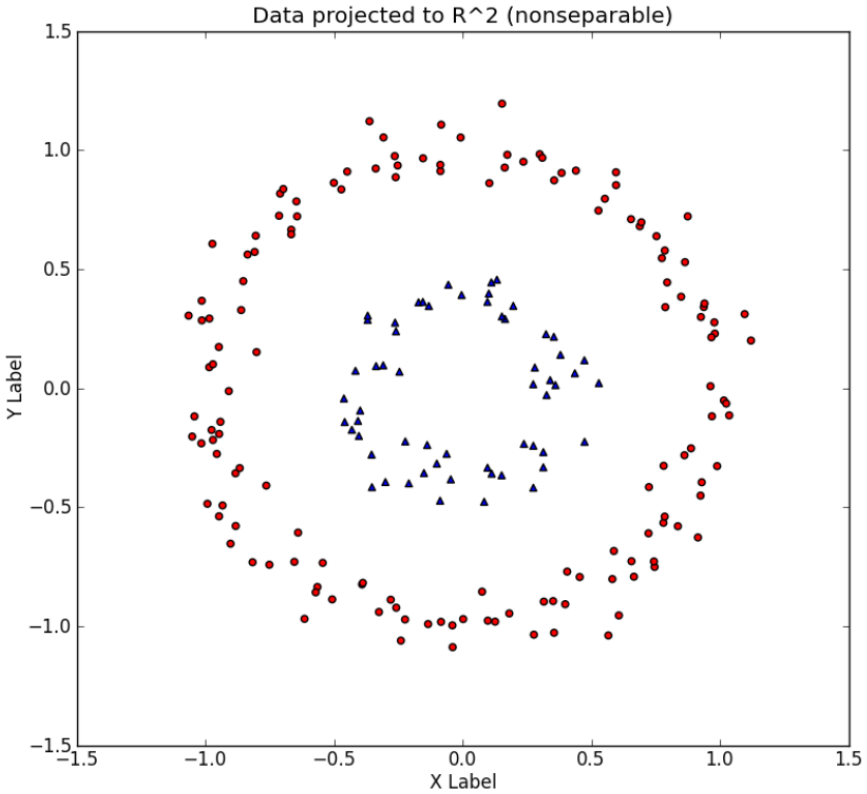
Plonger les données dans un espace de plus grande dimension :

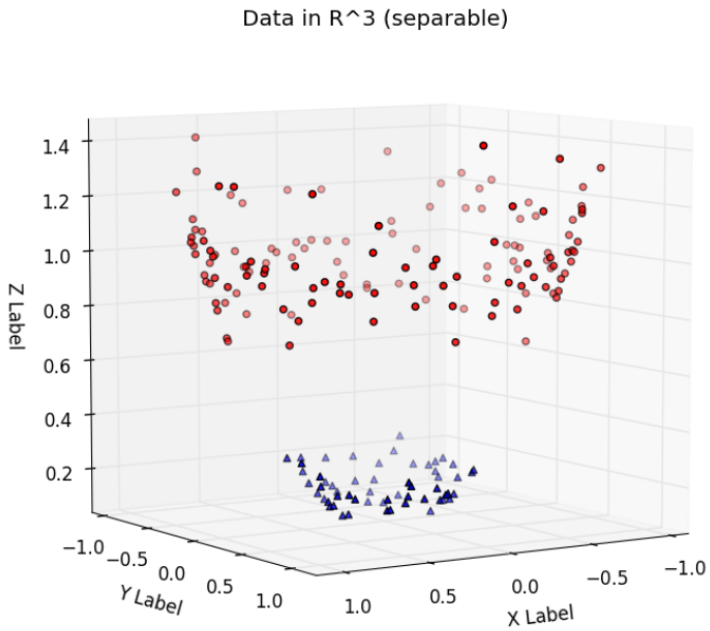
$$\Phi : X \longrightarrow \mathcal{F},$$

où les données **deviennent linéairement séparables**.

Idée générale

- La plupart des problèmes de classification relèvent de séparations non-linéaires.
- Mais l'espace des données peut toujours être plongé dans un espace de plus grande dimension dans lequel les données peuvent être linéairement séparées.





Feature space

Au lieu de chercher un hyperplan dans l'espace des entrées $\mathcal{X} = \mathbb{R}^d$, on passe d'abord dans un espace intermédiaire (feature space) de plus grande dimension :

$$\Phi : \mathcal{X} \longrightarrow \mathcal{F}, \qquad \mathbf{x} \longmapsto \Phi(\mathbf{x})$$

Exemple :
Soit $\mathbf{x} \in \mathbb{R}^2$, on considère la transformation :

$$\Phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3, \qquad (x[1], x[2]) \longmapsto \Phi(\mathbf{x}) = (x[1]^2, \sqrt{2}x[1]x[2], x[2]^2)$$

Astuce du noyau

La fonction Noyau ou Kernel représente le produit scalaire associé à l'espace de représentation intermédiaire. Pour une fonction Noyau (Kernel) $K(x_i, x_j)$, il existe un espace de Hilbert $\mathcal{F}$ et une fonction de transformation $\Phi(.)$ tels que :

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$$

- En pratique, l'astuce du noyau consiste à choisir une fonction Noyau, sans nécessairement caractériser l'espace $\mathcal{F}$ et la fonction $\Phi(.)$.
- Exercice : Déterminer la fonction Kernel associée à la transformation de l'exemple précédent.

Formulation du problème primal

Le problème primal du SVM avec marge douce et transformation de l'espace des données d'entrée est défini par

$$\min_{\omega,b,\xi\geq 0} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.c.} \quad \begin{cases} y_i(\omega\cdot\Phi(x_i)+b) \geq 1-\xi_i, & i=1,\cdots,n \\ \xi_i \geq 0, & i=1,\cdots,n \end{cases}$$

Remarque : le problème primal implique d'utiliser (et de connaître) la transformation $\Phi(\mathbf{x})$, ce qui n'est pas le cas du problème dual qui n'implique que la fonction Noyau (*Astuce du noyau*).

Formulation duale

Le problème dual du SVM avec marge douce et astuce du noyau (*kernel trick*) est défini par :

$$\alpha^* = \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{s.c.} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, \quad i = 1, \cdots, n \end{cases}$$

Solution optimale

Pour tout point $\mathbf{x} \in \mathcal{X}$, la fonction de décision optimale est :

$$h(\mathbf{x}) = \omega^* \cdot \Phi(\mathbf{x}) + b^* = \sum_{i \in \text{Supp}} \alpha_i^* y_i K(x_i, \mathbf{x})$$

et

$$b^* = y_j - \sum_{i \in \text{Supp}} \alpha_i^* y_i K(x_i, x_j) \quad \forall j \in \text{Supp}$$

Quelques noyaux usuels

- **Noyau linéaire** :

$$K(x_i, x_j) = \langle x_i, x_j \rangle$$

- **Noyau polynomial de degré** p :

$$K(x_i, x_j) = (c + \langle x_i, x_j \rangle)^p$$

- **Noyau Gaussien** :

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

- ...