Théorie et Algorithmes de l'Apprentissage Automatique

4 - Support Vector Machine

Simon BERNARD
simon.bernard@univ-rouen.fr

- SVM = *Support Vector Machine*, souvent traduit par Séparateur à Vaste Marge
- Initialement pour la classification binaire avec $y \in \{-1, 1\}$
- Séparateur linéaire :

$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b, \quad \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$$

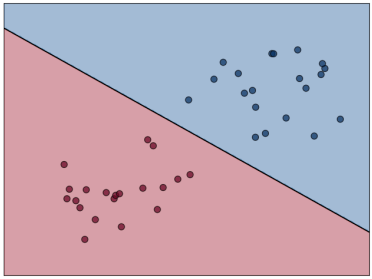
- Fonction de prédiction $sign(h(\mathbf{x}))$:

$$h(\mathbf{x}) < 0 \quad \text{on prédit -1 pour } \mathbf{x}$$
$$h(\mathbf{x}) > 0 \quad \text{on prédit 1 pour } \mathbf{x}$$

- Le séparateur linéaire est un hyperplan :

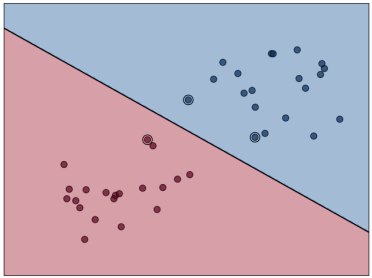
$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = 0$$

où $\mathbf{w}$ est le vecteur normal à l'hyperplan et b est le décalage par rapport à l'origine



- $h(\mathbf{x}) < 0 \rightarrow$ "en dessous" de l'hyperplan
- $h(\mathbf{x}) > 0 \rightarrow$ "au dessus" de l'hyperplan

- L'hyperplan "optimal" est celui qui maximise la distance des points à l'hyperplan :

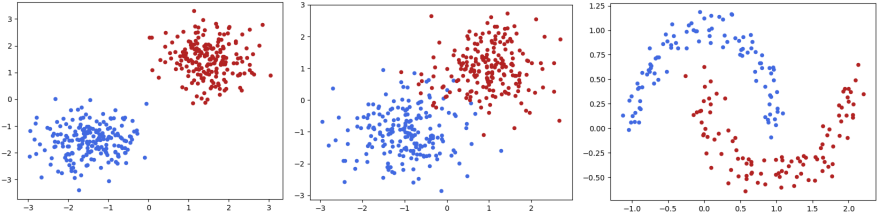


- On s'appuit sur les points les plus proches dans chaque classe
- Ces points sont appelés vecteurs supports
- La distance de ces points à l'hyperplan est appelée marge géométrique

Séparateur linéaire et marge

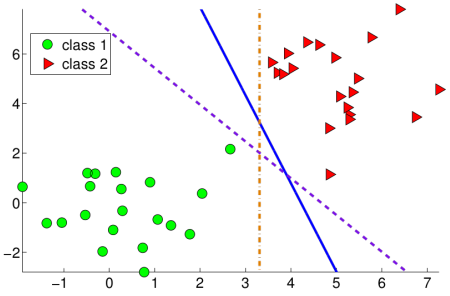
Definition

Les $\{(\mathbf{x}_i, y_i)\}$ sont linéairement séparables s'il existe un hyperplan qui permet de discriminer parfaitement l'ensemble des données. Dans le cas contraire, on parle d'exemples non linéairement séparables.



Dans un premier temps, considérons le cas linéairement séparable (e.g. figure de gauche)...

Objectif : trouver un hyperplan séparateur, qui sépare les points des classes 1 et 2



- Fonction de décision :

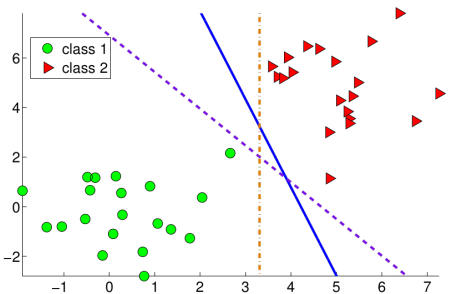
$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

$h(\mathbf{x}) = 0$ hyperplan séparateur

$h(\mathbf{x}) < 0$ classe 1

$h(\mathbf{x}) > 0$ classe 2

Objectif : trouver un hyperplan séparateur, qui sépare les points des classes 1 et 2



- Fonction de décision :

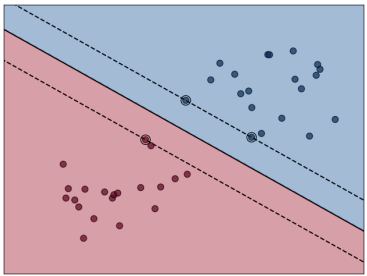
$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

- $\mathbf{w}$ est le vecteur normal de l'hyperplan
- b est le décalage par rapport à l'origine
- Plusieurs solutions possibles

- Problème : Quelle solution choisir ?
- Solution : Celle qui maximise la distance des points de chaque classe à l'hyperplan

Definition

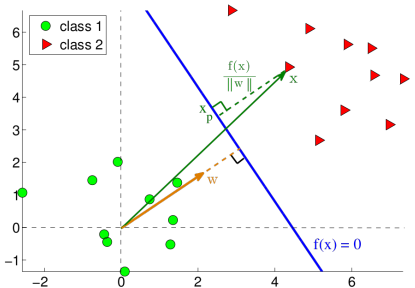
En apprentissage automatique, la marge d'un point **x** est définie comme la distance de **x** à la frontière de décision.



- Par extension, la marge géométrique de l'hyperplan est la distance entre les points les plus proches dans chaque classe (distance entre les droites en pointillés)
- Objectif : trouver l'hyperplan séparateur avec la plus grande marge

Distance d'un point $\mathbf{x} \in \mathbb{R}^d$ à l'hyperplan $H(\mathbf{v}, a) = \{\mathbf{z} \in \mathbb{R}^d \mid h(\mathbf{z}) = \mathbf{v}^\top \mathbf{z} + a = 0\}$:

$$d(\mathbf{x}, H) = \frac{|\mathbf{v}^\top \mathbf{x} + a|}{\|\mathbf{v}\|} = \frac{|h(\mathbf{x})|}{\|\mathbf{v}\|}$$



Soit $\mathbf{x}_p$ la projection orthogonale de $\mathbf{x}$ sur H.
On cherche d tel que :

$$\mathbf{x} = \mathbf{x}_p + d\frac{\mathbf{v}}{\|\mathbf{v}\|}$$

$$d\frac{\mathbf{v}}{\|\mathbf{v}\|} = \mathbf{x} - \mathbf{x}_p$$

$$d\frac{\|\mathbf{v}\|^2}{\|\mathbf{v}\|} = \mathbf{v}^\top \mathbf{x} + a - \underbrace{(\mathbf{v}^\top \mathbf{x}_p + a)}_{=0}$$

$$d = \frac{|\mathbf{v}^\top \mathbf{x} + a|}{\|\mathbf{v}\|}$$

- Trouver l'hyperplan qui maximise la marge, revient à résoudre

$$\max_{\mathbf{v},a} \quad \underbrace{\min_{i=1,\dots,n} d(\mathbf{x}_i, H)}_{m}$$

- Ce problème peut s'écrire comme un problème d'optimisation sous contraintes :

$$\begin{aligned} \max_{\mathbf{v},a} \quad & m \\ \text{s.c.} \quad & \frac{|\mathbf{v}^\top \mathbf{x}_i + a|}{\|\mathbf{v}\|} \geq m, \quad i = 1, \dots, n \end{aligned}$$

- Problème : plusieurs $\mathbf{v}$ pour un même hyperplan donc plusieurs solutions

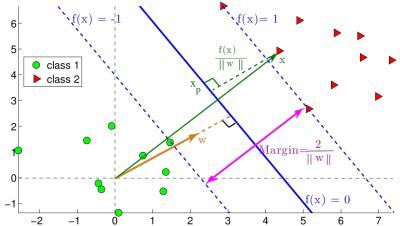
- Solution : on pose $\mathbf{w} = \frac{\mathbf{v}}{m\|\mathbf{v}\|}$ et $b = \frac{a}{m\|\mathbf{v}\|}$:

$$\mathbf{w} = \frac{\mathbf{v}}{m\|\mathbf{v}\|} \Rightarrow \|\mathbf{w}\| = \frac{\|\mathbf{v}\|}{m\|\mathbf{v}\|} = \frac{1}{m} \Rightarrow m = \frac{1}{\|\mathbf{w}\|}$$

$$\frac{|\mathbf{v}^\top \mathbf{x}_i + a|}{\|\mathbf{v}\|} \geq m \Rightarrow \frac{|\mathbf{v}^\top \mathbf{x}_i + a|}{m\|\mathbf{v}\|} \geq 1 \Rightarrow |\mathbf{w}^\top \mathbf{x}_i + b| \geq 1 \Rightarrow y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1$$

car $\mathbf{w}^\top \mathbf{x}_i + b \geq 0$ quand $y_i = 1$ et $\mathbf{w}^\top \mathbf{x}_i + b \leq 0$ quand $y_i = -1$

- Le changement de variables revient à imposer que $|h(\mathbf{x}_i)| \geq 1, i = 1, \ldots, n$

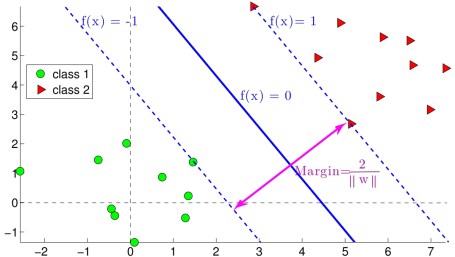


- $h(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b \geq 1$ quand $y_i = 1$
- $h(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b \leq -1$ quand $y_i = -1$
- Dans ce cas, la marge géométrique est $\frac{2}{\|\mathbf{w}\|}$

- Un classifieur SVM est définit par $D(\mathbf{x}) = signe(h(\mathbf{x}))$ avec $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ tel que :

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{maximisation de la marge}$$

$$\text{s.c.} \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1, i = 1, \ldots, n \qquad \text{tous les points sont bien classés}$$



Résolution du problème d'optimisation SVM

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{z}^\top \mathbf{A}\mathbf{z} - \mathbf{d}^\top \mathbf{z} \\ \text{s.c.} \quad & \mathbf{B}\mathbf{z} \leq \mathbf{e} \end{aligned}$$

avec $\mathbf{z} = (\mathbf{w}, b)^\top \in \mathbb{R}^{d+1}$, $\mathbf{d} = (0, \ldots, 0)^\top \in \mathbb{R}^{d+1}$, $\mathbf{A} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix}$, $\mathbf{I}$ est la matrice identité de $\mathbb{R}^d$, $\mathbf{B} = -[diag(\mathbf{y})\mathbf{X}, \mathbf{y}]$ et $\mathbf{e} = -(1, \ldots, 1)^\top$

- Ce problème est convexe (car $\mathbf{A}$ est définie positive), donc il existe une solution unique
- La résolution peut se faire avec des méthodes dédiées
- Cependant, le problème peut être transformé pour rendre la résolution plus efficace...

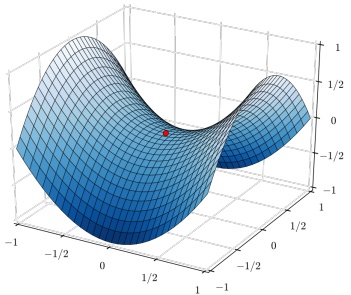
- Problème primal de SVM :

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.c.} \quad & y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1, i = 1, \ldots, n \end{aligned}$$

- On introduit les multiplicateurs de Lagrange $\alpha_i \geq 0$ associés aux n contraintes d'inégalités
- La fonction résultante est appelée Lagrangien :

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\right)$$

- La solution $(\mathbf{w}^*, b^*)$ est le point-selle du Lagrangien : $L(\mathbf{w}, b, \boldsymbol{\alpha}^*)$ est minimal en $(\mathbf{w}^*, b^*)$ et $L(\mathbf{w}^*, b^*, \boldsymbol{\alpha})$ est maximal en $\boldsymbol{\alpha}^*$.



En rouge : point-selle de $(x, y) \to x^2 - y^2$

Un point-selle d'une fonction $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ est un point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ tel que

- $y \to f(x^*, y)$ atteint un maximum en y^*
- $x \to f(x, y^*)$ atteint un minimum en x^*

$$\max_{\boldsymbol{\alpha}} \quad \min_{\mathbf{w},b} \quad L(\mathbf{w}, b, \boldsymbol{\alpha})$$

- En calculant les gradients :

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \qquad \frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = 0$$

On obtient les conditions d'optimalité (stationnarité) :

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

- Le problème dual s'écrit d'abord comme :

$$\begin{aligned} \max_{\mathbf{w},b,\boldsymbol{\alpha}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \right) \\ \text{s.c.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{et} \quad \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \end{aligned}$$

- Mais on peut simplifier le Lagrangien avec $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \Big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \Big)$$

$$L(\boldsymbol{\alpha}) = \underbrace{\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i}_{\mathbf{w}^\top \mathbf{w}} - \sum_{i=1}^{n} \alpha_i y_i \underbrace{\sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j^\top \mathbf{x}_i}_{\mathbf{w}^\top} - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0} + \sum_{i=1}^{n} \alpha_i$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i$$

- Problème dual de SVM :

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i \\ \text{s.c.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

- Pour les problèmes convexes, le dual a la même solution que le primal
- Résoudre le dual permet de trouver les n paramètres $\alpha_i, i = 1, \ldots, n$

- On obtient deux types de α_i :
  - si $y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1$ alors $\alpha_i = 0$
  - si $y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$ alors $\alpha_i > 0$
- Solution : $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$
- Donc $\mathbf{w}$ n'est défini que par les points qui vérifient $y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$
- Ces points sont les vecteurs supports

- Calcul de $\mathbf{w}$:
  - Utiliser les données $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ pour résoudre le dual
    $\rightarrow$ on obtient les paramètres $\{\alpha_i, i = 1, \ldots, n\}$
  - En déduire la solution $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$

- Calcul de b :
  - Les $\alpha_i > 0$ correspondent aux points supports qui vérifient :

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$$

  - En déduire la valeur de b

- La fonction de décision :

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

Primal

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.c.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, n$$

- Apprentissage de $d + 1$ paramètres
- Donnent l'influence des caractéristiques
- Programme quadratique classique
- Parfait quand $d \ll n$
- Coûteux si d est grand (calcul de $\mathbf{w}^\top \mathbf{x}$)

Dual

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i$$
$$\text{s.c.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, n$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

- Apprentissage de n paramètres
- Donnent l'influence des exemples
- Plus simple à résoudre
- Utiliser quand $d > n$
- Permet l'utilisation de fonction noyaux (nous y reviendrons)

Problèmes non linéairement séparables

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{maximisation de la marge}$$

$$\text{s.c.} \quad y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1, i = 1, \ldots, n \qquad \text{tous les points sont bien classés}$$

- Relâcher les contraintes car tous les points ne peuvent pas être correctement classés
- On introduit des variables "d'erreur" ξ_i pour chaque point :

$$y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i$$

où $\xi_i \geq 0, \forall i = 1, \ldots, n$

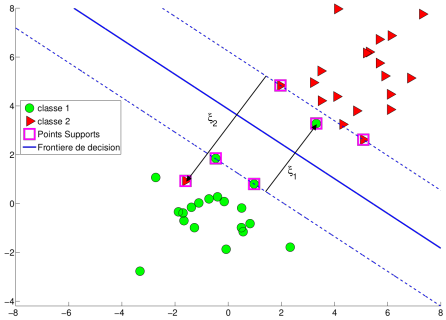
- Deux situations :

$$\text{pas d'erreur} \quad y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 \Rightarrow \qquad \xi_i = 0$$

$$\text{erreur} \quad y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) < 1 \Rightarrow \qquad \xi_i = 1 - y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) > 0$$

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{maximisation de la marge}$$

$$\text{s.c.} \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1, i = 1, \ldots, n \qquad \text{tous les points sont bien classés}$$



$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{maximisation de la marge}$$

$$\text{s.c.} \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1, i = 1, \ldots, n \qquad \text{tous les points sont bien classés}$$

- Inclure la somme des "erreurs" dans le problème SVM

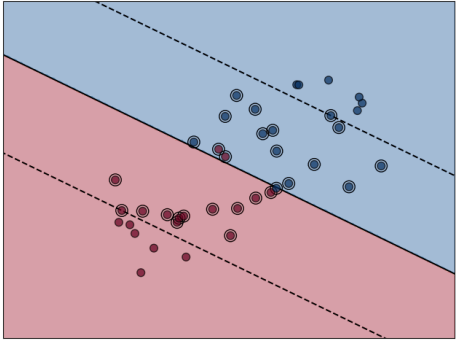
$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

avec $C > 0$

- C est appelé paramètre de régularisation (compromis entre erreur et marge)
- C'est un hyper-paramètre à fixer par l'utilisateur

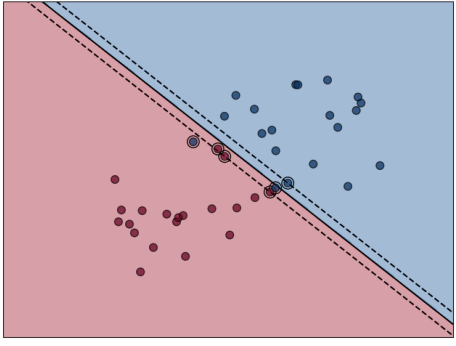
Illustration

L'hyper-paramètre C représente l'influence maximale de chaque exemple



$C = 0.01$

Quand C petit $\rightarrow$ marge grande



$C = 1000$

Quand C grand $\rightarrow$ marge petite

- Nouveau problème primal :

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.c.} \quad & y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i, && i = 1,\dots,n \\ & \xi_i \geq 0, && i = 1,\dots,n \end{aligned}$$

- Lagrangien :

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{n}\beta_i\xi_i$$

avec $\alpha_i \geq 0, \beta_i \geq 0, \forall i = 1,\dots,n$.

- Conditions d'optimalité :

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial b} = 0 \quad \frac{\partial L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \mathbf{w}} = 0 \quad \frac{\partial L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \xi_i} = 0$$

- Ce qui donne

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad C - \alpha_i - \beta_i = 0, \forall i = 1, \ldots, n$$

- En ré-injectant $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ dans la Lagrangien, on obtient à nouveau

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i$$

- La fonction objective est la même, mais des contraintes s'ajoutent...

- Or, comme $\beta_i \geq 0$, la troisième condition d'optimalité implique que :

$$0 \leq \alpha_i \leq C$$

- Donc le nouveau problème dual est

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i \\ \text{s.c.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

SVM en pratique

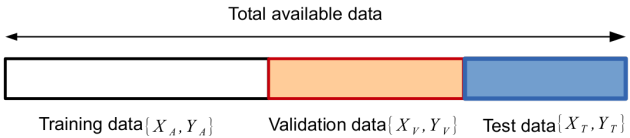
- Entrées :

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\}, i = 1, \ldots, n\}$$

- Procédure :
  1. Centrer et réduire les données $\{\mathbf{x}_i, i = 1, \ldots, n\} \rightarrow \{\mathbf{x}_i = \mathbf{\Sigma}^{-1}(\mathbf{x}_i - \overline{x}), i = 1, \ldots, n\}$
  2. Fixer le paramètre $C > 0$ du SVM
  3. Utiliser un solveur pour résoudre le problème dual et obtenir les $\alpha_i \neq 0$, les points supports $\mathbf{x}_i$ correspondants et le biais b
  4. En déduire la fonction de décision $h(\mathbf{x}) = \sum_{\mathbf{x}_i \in SV} \alpha_i y_i \mathbf{x}_i \mathbf{x} + b$
  5. Evaluer l'erreur en généralisation du SVM obtenu. Recommencer en 2. si elle n'est pas satisfaisante

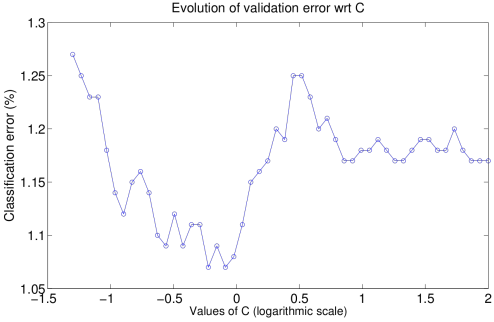
- *LibSVM* : `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
  C, avec beaucoup d'interface dans d'autres languages. Beaucoup de librairies s'appuient sur ce solver
- *Scikit-learn* : `https://scikit-learn.org/stable/modules/svm.html`
- *PyML* : `http://pyml.sourceforge.net/#`
  Librairie Python spécialisée dans les SVM et les méthodes à noyaux
- *Shogun* : `http://www.shogun-toolbox.org/page/Events/gsoc2011_ideas`
  Librairie de ML en C++, complète et performante

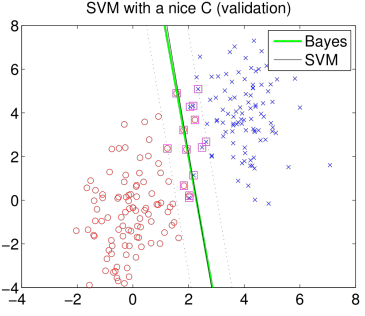
- On commence par séparer les données en trois sous-ensembles :
  - Ensemble d'apprentissage : pour calculer $\mathbf{w}$ et b
  - Ensemble de validation : pour évaluer l'erreur pour différents C
  - Ensemble de test : pour évaluer le modèle final



1. Pour différentes valeurs de C :
   - $(\mathbf{w}, b) \leftarrow \mathit{TrainLinearSVM}(X_a, Y_a, C, \mathit{options})$
   - $\mathit{error} \leftarrow \mathit{EvaluateError}(X_v, Y_v, \mathbf{w}, b)$
2. $C \leftarrow \arg\min \mathit{error}$

- Les valeurs de C Choisies sur une échelle logarithmique
- Pour chaque C, on apprend un SVM et on calcule son erreur en validation
- Le minimum de la courbe d'erreur correspond à la "meilleure" valeur de C sur ce problème
- Le SVM finale correspondant est sur la figure de droite





Pour une problème à $K > 2$ classes, deux approches populaires

- Approche *One-vs-All* (OVA) :
  - K SVM avec une classe comme classe positive et toutes les autres comme classe négative
  - Classe prédite : le plus grande score en tant que classe positive (*winner takes all*)

$$D(\mathbf{x}) = \underset{k=1,K}{\arg\max}\, \mathbf{w}_k^\top \mathbf{x} + b_k$$

- Approche *One-vs-One* (OVO) :
  - $\frac{K(K-1)}{2}$ SVM : un par paire de classes
  - Classe prédite : vote à la majorité ou estimation de la probabilité *a posteriori* maximale

- Construction d'un hyperplan optimal au sens de la maximisation de la marge
- Une analyse théorique poussée montre que maximiser la marge équivaut à minimiser une borne sur l'erreur de généralisation.
- Le cas non linéaire (où on cherche une fonction de décision non-linéaire) peut être traité grâce aux noyaux (prochain chapitre)
- Généralisation possible au cas où on a plusieurs classes (*one-vs-all*, *one-vs-one*,...)
- Algorithme de classification très utilisé en pratique...