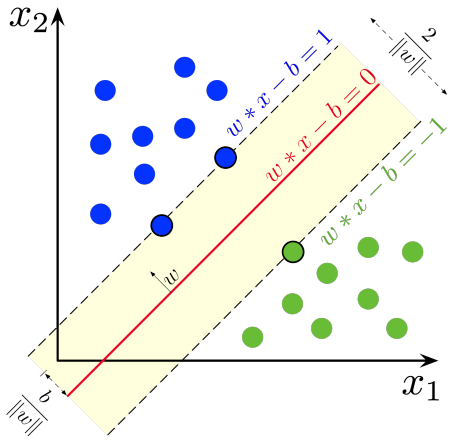
SVM Linéaires

Objectif de la SVM

Une SVM cherche un **hyperplan optimal** séparant deux classe de données en maximisant la **marge** (distance entre l'hyperplan et les points les plus proches, appelés *vecteurs de support*). Avec gestion des erreurs, une *marge douce* est introduite via un coefficient C, pour tolérer des erreurs de classification et améliorer la robustesse.
Si on veut utiliser un modèle de SVM sans erreur (qui n'a pas toujours de solutions) on supprime simplement les ξ_i dans toutes les équations (C disparait aussi).

Définition de la Marge

Pour une SVM linéaire, l'hyperplan est défini par $w^T x + b = 0$, avec w le **vecteur de poids**, b le **biais**, et x un point. Les hyperplans marginaux sont $w^T x + b = \pm 1$. La distance entre les deux est $\frac{2}{\|w\|}$, soit la **marge**. L'objectif est de maximiser la marge, donc de minimiser $\|w\|$.



Gestion des Erreurs : Marge Douce

Si on ne considère pas les erreurs, on veut $(w^T x_i + b) \geq 1$ si $y = 1$ et $(w^T x_i + b) \leq 1$ si $y = -1$, ce qui équivaut à $y_i(w^T x_i + b) \geq 1$. Pour les points mal classés, on introduit des variables de **relaxation** $\xi_i \geq 0$:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n$$

où ξ_i représente le degré de violation de la marge pour le point x_i. La somme des ξ_i est pénalisée dans la fonction objectif pour limiter les erreurs.

Problème d'Optimisation Primal avec C

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

sous contraintes : $y_i(w^T x_i + b) \geq 1 - \xi_i$ et $\xi_i \geq 0, \quad \forall i = 1, \ldots, n$. Le **paramètre** C contrôle le compromis entre la marge large et les erreurs de classification (plus C est grand, plus les erreurs sont pénalisées et plus la marge est petite).

Lagrangien et Dualité

En introduisant des **multiplicateurs de Lagrange** $\lambda_i \geq 0$ et $\mu_i \geq 0$ pour chaque contrainte, le Lagrangien devient :

$$L(w, b, \xi, \lambda, \mu) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i \left(y_i(w^T x_i + b) - 1 + \xi_i \right) - \sum_{i=1}^{n} \mu_i \xi_i$$

Hugo Tondenier

Conditions d'Optimalité (KKT)

En annulant les dérivées par rapport à w, b, et ξ (c'est a dire en miniminsant le lagrangien), on obtient :

$$w = \sum_{i=1}^{n} \lambda_i y_i x_i$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

$$\lambda_i = C - \mu_i$$

avec $\xi_i > 0$ si $\lambda_i = C$, indiquant que le point x_i est mal classé ou se trouve dans la marge.

Passage au Problème Dual

1. Substituons $w = \sum_{j=1}^{n} \lambda_j y_j x_j$ dans L :

$$L = \frac{1}{2} \left(\sum_{j=1}^{n} \lambda_j y_j x_j \right)^T \left(\sum_{k=1}^{n} \lambda_k y_k x_k \right) + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i \left(y_i \left(\sum_{j=1}^{n} \lambda_j y_j x_j^T x_i + b \right) - 1 + \xi_i \right) - \sum_{i=1}^{n} (C - \lambda_i) \xi_i$$

2. Les termes en ξ_i se simplifient :

$$C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i \xi_i - \sum_{i=1}^{n} (C - \lambda_i) \xi_i = 0$$

3. Le Lagrangien simplifié devient :

$$L = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \lambda_j \lambda_k y_j y_k x_j^T x_k - \sum_{i=1}^{n} \lambda_i y_i \left(\sum_{j=1}^{n} \lambda_j y_j x_j^T x_i + b \right) + \sum_{i=1}^{n} \lambda_i$$

4. En utilisant la condition $\sum_{i=1}^{n} \lambda_i y_i = 0$, le terme en b disparaît. Le Lagrangien final est :

$$L = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

Problème Dual avec Marge Douce

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j (x_i^T x_j)$$

sous contraintes (vient des conditions KKT) : $0 \leq \lambda_i \leq C$ (car C et λ_i positifs) et $\sum_{i=1}^{n} \lambda_i y_i = 0$.

Résolution du dual $\Rightarrow$ valeurs optimales des λ_i et identification des *vecteurs de support* (points pour lesquels $\lambda_i > 0$).

Modèle Final

Avec les λ_i, on calcule w par :

$$w = \sum_{i=1}^{n} \lambda_i y_i x_i$$

et le biais b est obtenu en utilisant un vecteur de support x_k tel que $0 < \lambda_k < C$:

$$b = y_k - w^T x_k$$

Prédiction avec la SVM

Pour prédire la classe d'un nouveau point x, on utilise :

$$f(x) = \text{sign}\left(w^T x + b\right) = \text{sign}\left(\sum_{i=1}^{n} \lambda_i y_i (x_i^T x) + b \right)$$

Seuls les vecteurs de support (pour lesquels $\lambda_i > 0$) influencent la décision.

Hugo Tondenier