École nationale supérieure des Mines de Nancy

Rapport de projet 3A

Pierre Gauthier

Algorithme d'apprentissage en chimie quantique et application au screening (sélection) de cellules photovoltaïques

Laboratoire : Institut Élie Cartan

Tuteurs : Dario Rocca et Marianne Clausel

21 Novembre 2018

Table des matières

Introduction :

Ce projet est conduit dans un cadre pédagogique en tant que projet de troisième année à l'Ecole des Mines de Nancy. Il suit la publication scientifique de Mathias Rupp "Machine Learning for Quantum Mechanics in a Nutshell". Dans cette publication Mathias Rupp propose d'allier la mécanique quantique aux méthodes de machines learning pour faire de la prédiction à partir de données et ainsi dépasser les problèmes en terme de puissance de calcul du problème à N-corps. Il vise ainsi à prédire l'énergie d'atomisation de molécules à partir d'un set de données d'entrainement, en utilisant des méthode de linéaire, notamment la régression à vecteur supports (SVM). Le premier objectif du projet est de reproduire les résultats de cette études, et d'explorer des variations dans les paramètres sur l'erreur finale de prédiction. Nous pourrons également dépasser le travail réaliser dans l'étude en travaillant sur des données avec de nouveaux descripteurs qui prennent en compte les propriétés des groupes chimiques des molécules.
Nous allons dans une première partie présenterer les méthodes à vecteur support avec l'astuce des noyaux.

Chapitre 1

Support Vector Machines methods (SVM)

Le probléme SVM vise à séparer les données $(x_i,\ y_i)_{1\leqslant i\leqslant N},\ x_i \in \mathbb{R},\ y_i \in \{-1,\ 1\}$ en deux classes $+1$ et -1 à l'aide de la fonction $f(x) = w \cdot x + b$ $(b \in \mathbb{R},\ w \in \mathbb{R}^d)$ telle que $f(x) > 0 \Rightarrow x \in C_{+1}$, et $f(x) < 0 \Rightarrow x \in C_{-1}$

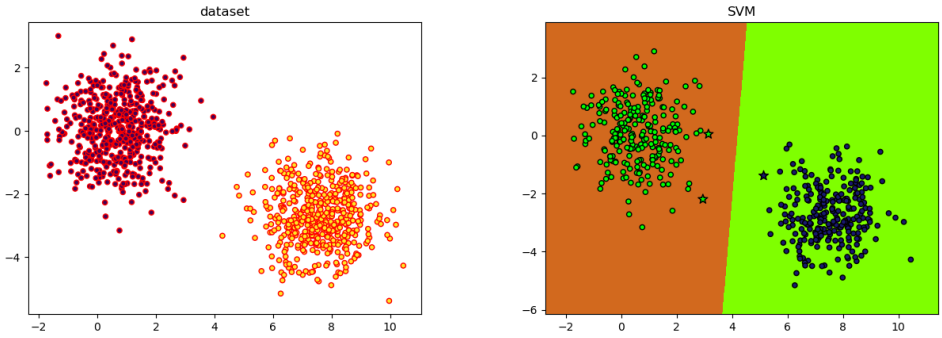

FIGURE 1.1 – sparation de données générées make_blobs du package dataset et séparation de l'espace en deux classe par la méthode des vecteurs support à l'aide de la fonction SVC du package sklearn. Les étoiles sont les vecteurs supports.

Nous voulons trouver l'hyperplan qui sépare le mieux nos données parmi tous ceux compatibles.

Pour juger la qualité d'un hyperplan en tant que séparateur on utilise la distance entre les exemples d'apprentissage et ce séparateur. Plus précisément, la « marge » d'un problème d'apprentissage est définie comme la distance entre le plus proche exemple d'apprentissage et l'hyperplan de séparation.

Pour un hyperplan H, On a :

$$\text{Marge}(H) = \min_{x_i}\ d(x_i,\ H)$$

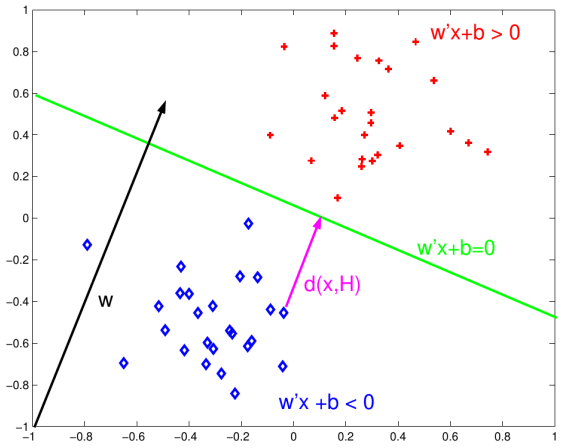

FIGURE 1.2 – Le séparateur idéal correspond intuitivement à l'hyperplan qui passe « au milieu » entre les données sans préférence pour une classe ou une autre. C'est le séparateur de marge maximale.[Cours Cnam RCP209]

Sous l'hypothèse qu'il exixte un hyperplan qui sépare nos données, trouver l'hyperplan qui maximise la marge revient à résoudre le problème suivant :

$$\begin{cases} \arg\min_{w,b}\ \dfrac{1}{2}||w||^2 \\ \forall\ 1 \leqslant i \leqslant N,\ y_i(w \cdot x_i + b) \geqslant 1 \end{cases}$$

On utilise le lagrangien des conditions de Karush, Kuhn et Tucker, qui s'exprime sous la forme suivante :

$$L(w, b, \lambda_i) = \frac{1}{2}||w||^2 - \sum \lambda_i (y_i(w \cdot x_i + b) - 1)$$

On recherche donc le λ qui maximise

$$\max\ L(\lambda) = \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i \cdot x_j$$

$$\text{sous contraintes } \lambda_i \geqslant 0 \text{ et } \sum_i \lambda_i y_i = 0$$

On utilise alors l'astuce du noyaux qui consiste à remplacer le produit scalaire $x \cdot y$ par un noyaux reproduisant $K(x, y) = \phi(x) \cdot \phi(y)$, $K : \xi \to \mathbb{R}$, $\phi : \xi \to \mathbb{R}$. Le théorème de Mercer assure l'existence d'une telle décomposition du noyaux K

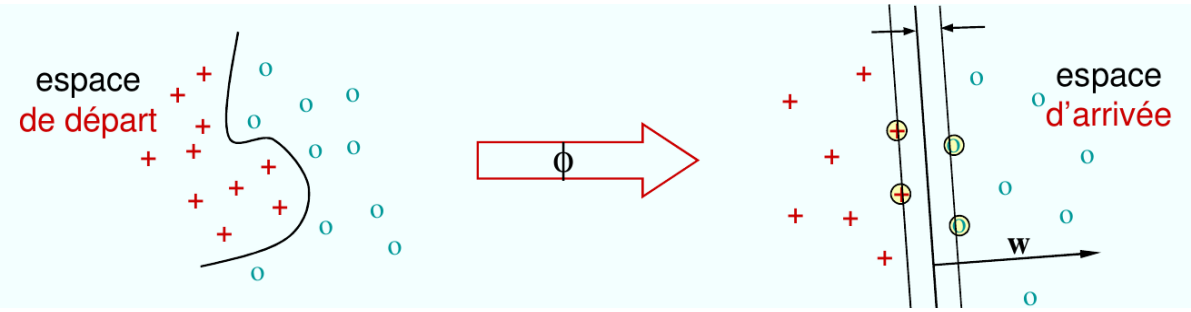

FIGURE 1.3 – Astuce à noyaux : projeter les données dans un espace de dimension beaucoup plus grande, où elles deviennent séparables linéairement.[Cours Cnam RCP209]

Chapitre 2

Experimentations