Soutenance projet de troisième année : Algorithme d'apprentissage en chimie quantique et application au screening (sélection) de cellules photovoltaïques

Etudiant : Pierre Gauthier
Tuteurs : Jérémie Unterberg, Marianne Clausel, Dario Rocca

5 Février 2019

Introduction

- Travail sur la publication de Mathias Rupp *Machine Learning for Quantum Mechanics in a Nutshell*, s'inscrit dans le *Harvard Clean Energie Project*
- Couplage Physique Quantique / Apprentissage automatique

1. Description des données
2. Outils d'apprentissage automatique
   - Support Vector Machines methods (SVM)
   - Astuce du noyau
   - Ridge Regression
3. Expérimentations
   - Base d'entrainement et de Test du modèle
   - Fonctions d'erreurs utilisées
   - Recherche hyparamètres optimaux
4. Perspectives
5. Conclusion

Description des données

- Dataset de molécules en .xyz :

Nombre d'atome			
numéro de la molécule	énergie d'atomisation		
atome 1	x(1)	y(1)	z(1)
atome 2	x(2)	y(2)	z(2)
.	.	.	.
.	.	.	.
.	.	.	.
atome n	x(n)	y(n)	z(n)

Description des données

- Utilisation des Matrices de Coulomb (Z numéro atomique) :

$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4} & i = j \\ \dfrac{Z_i Z_j}{||R_i - R_j||_2} & i \neq j \end{cases}$$

- Matrice sysmétrique, de taille $23 \times 23 \rightarrow$ stockage dans vecteur taille $\frac{23 \times (23+1)}{2} = 276$.
- La matrice varie avec permutation des atomes $\rightarrow$ Tri des lignes par norme décroissante

Description des données

Exemple simple de matrice Coulombs pour C−H et H−C−H :

	C	H		
C	a_1	b_1	0	0
H	b_1	a_2	0	0
	0	0	0	0
	0	0	0	0

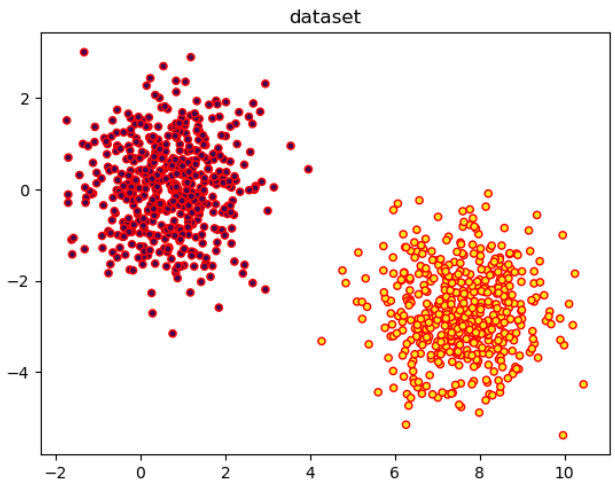
$\rightarrow$ (a_1, b_1, 0, 0, a_2, 0, 0, 0, 0, 0)

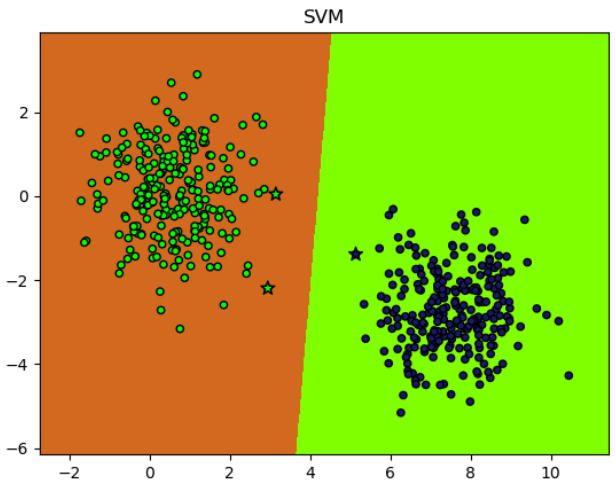
	C	H	H	
C	a_1	b_1	c_1	0
H	b_1	a_2	c_2	0
H	c_1	c_2	a_3	0
	0	0	0	0

$\rightarrow$ (a_1, b_1, c_1, 0, b_1, a_2, c_2,0 ,a_3, 0)

Support Vector Machines methods (SVM)

Nous allons commencer par présenter le problème de classification avant la régression

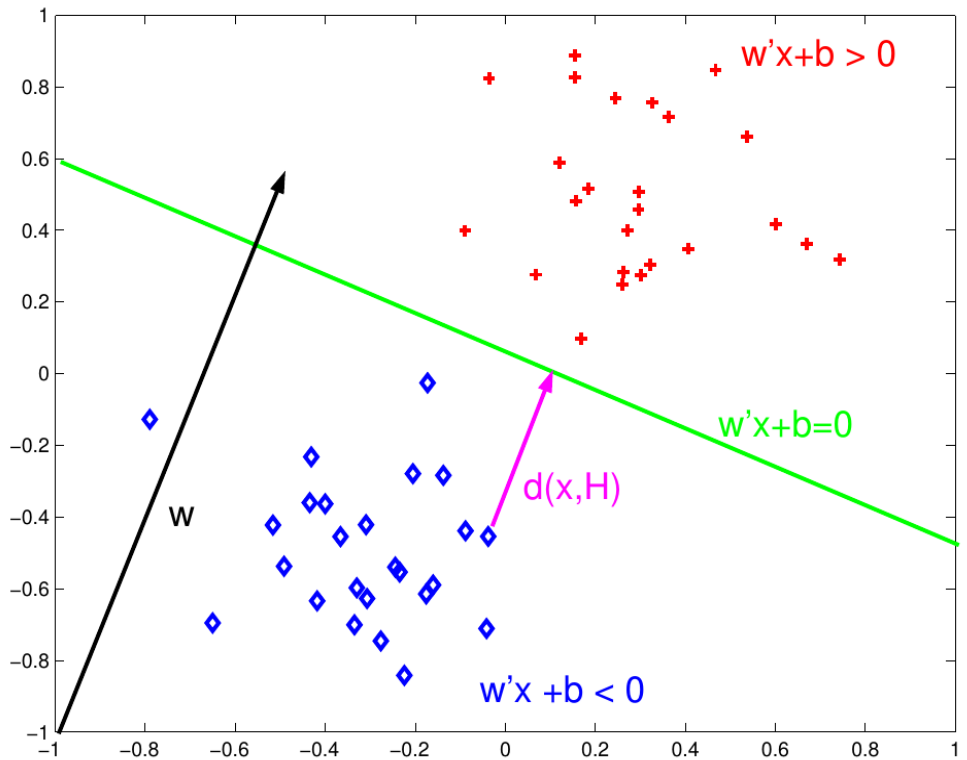




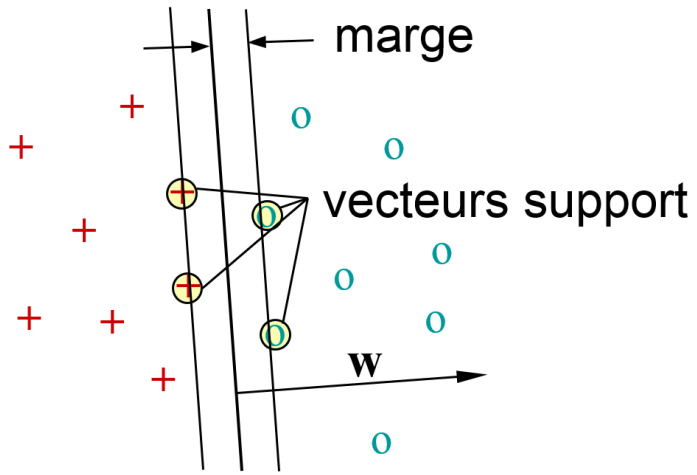
Support Vector Machines methods (SVM)

Hyperplan de séparation

$$H = \{x | w^T x + b = 0\} \quad \text{Marge}(H) = \min_{x_i} \; d(x_i, \; H)$$



Support Vector Machines methods (SVM)



$$2 \times \text{Marge} = 2 \times d(x,\ H) = \frac{|w^T x_{vs} + b|}{||w||}$$

$$|w^T x_{vs} + b| = 1 \quad \rightarrow \quad \text{Marge} = \frac{2}{||w||}$$

Support Vector Machines methods (SVM)

On en arrive au problème de minimisation suivant :

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

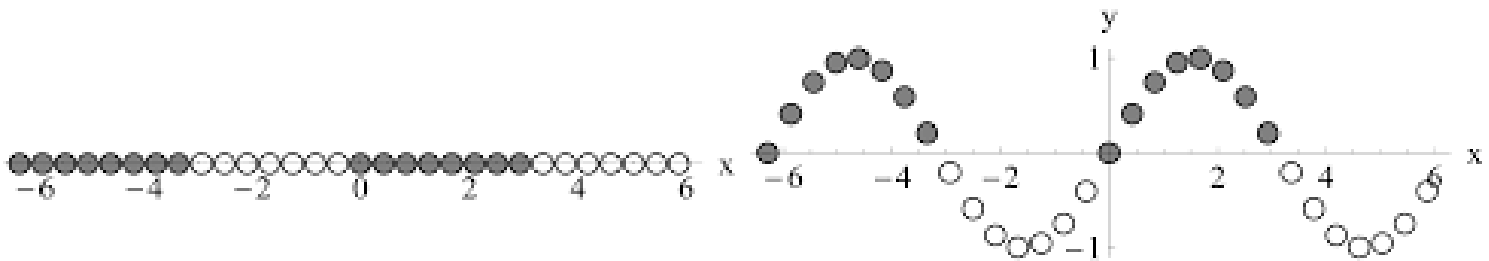
On formule le problème dual en introduisant le Laplacien :

$$\begin{cases} \max L(\lambda) = \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i \cdot x_j \\ \lambda_i \geqslant 0 \\ \sum_i \lambda_i y_i = 0 \end{cases}$$

$$f^*(x) = \sum_{i=1}^{n} \lambda_i^* y_i x_i^T x + b^*$$

Astuce du noyau

Augmentation de la dimension de l'espace pour rendre les données linéairement séparables



Linear inseparability in input space — Linear separability in transformed space

Astuce du noyau

$$f(x) = \sum_{i=1}^{n} \lambda_i y_i x_i^T x + b \rightarrow \sum_{i=1}^{n} \lambda_i y_i K(x_i, x) + b$$

Si K doit vérifier les conditions :

- K est continue symétrique
- $K(x_i, x_j)_{\ 1 \leqslant i,j \leqslant N}$ est une matrice définie positive

Alors il existe $\phi : \xi \rightarrow H$ telle que $K(x, y) = \langle \phi(x), \phi(y) \rangle$.

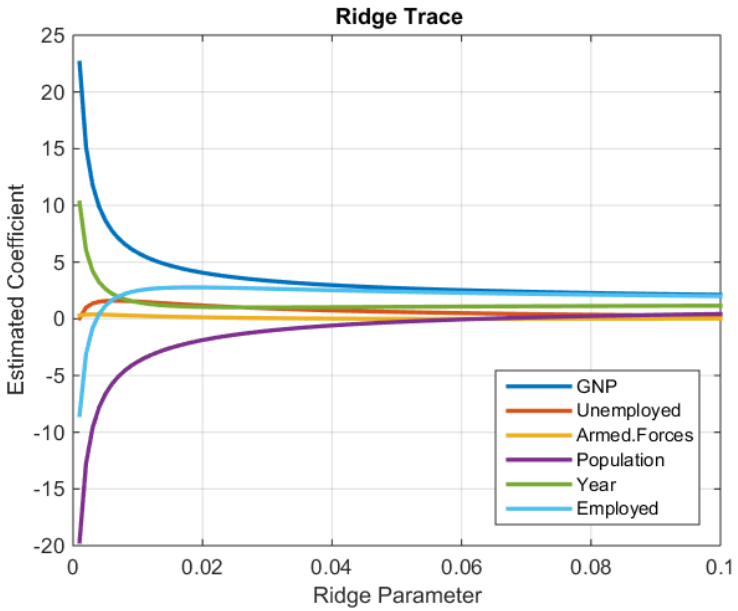
Ridge Régression

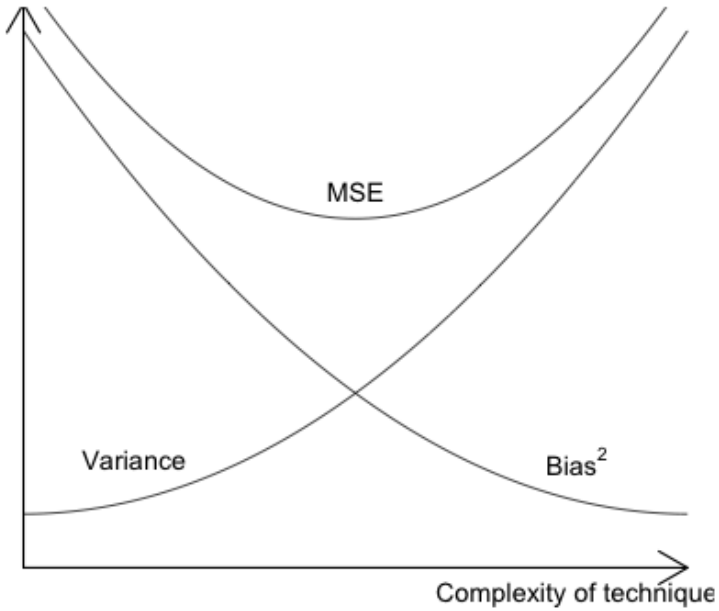
On ajoute un terme de pénalisation $|.|_2^2$ au problème de régression classique

$$\arg\min_{w \in \mathbb{R}^n} \sum (f(x_i) - y_i)^2 \quad \rightarrow \quad \arg\min_{w \in \mathbb{R}^n} \sum (f(x_i) - y_i)^2 + \lambda ||w||_2^2$$

$$w = (X^T X)^{-1} X^T y \quad \rightarrow \quad w^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Ridge Régression





Ridge Régression

Applications aux méthodes à noyaux avec $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$

$$\arg\min_{\alpha \in \mathbb{R}^n} \sum (f(x_i) - y_i)^2 + \lambda ||f||_H^2$$

$$\Leftrightarrow \arg\min_{\alpha \in \mathbb{R}^n} < K\alpha - y, K\alpha - y > + \lambda \alpha^T K \alpha$$

$$\Rightarrow \alpha = (K + \lambda I)^{-1} y$$

Avec $K \in \mathbb{R}^{n \times n}$ est la matrice du noyau $\quad K_{i,j} = K(x_i, x_j)$

Base d'entrainement et de Test du modèle

On séparer le dataset d'environ 7000 molécules en

- un set d'entrainement de 900 molécules $\rightarrow$ recherche des hyperparamètres (λ, paramètres des noyaux)
- un hold-out set de 100 molécules pour vérifier qu'il n'y a pas de sur-apprentissage sur le set d'entrainement
- Un set de prédiction avec reste des molécules.

Fonctions d'erreurs utilisées

- $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=}^{n}(y_i - f(x_i))^2}$
- $\text{MAE} = \frac{1}{n}\sum_{i=}^{n}|y_i - f(x_i)|$
- $(1 - R^2)\sum_{i=1}^{n}(\overline{y} - y_i)^2 = \sum_{i=}^{n}(y_i - f(x_i))^2$

Recherche hyparamètres optimaux

Présentation des noyaux utilisés :

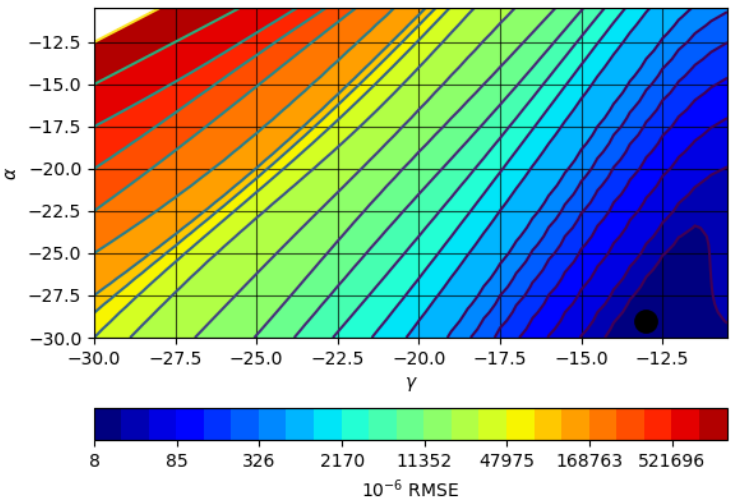
Noyau Gaussien

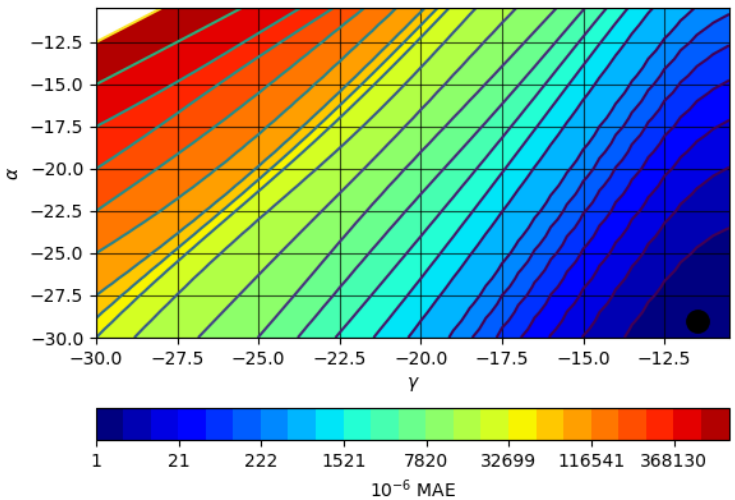
$$K(x,z) = \exp -\frac{||x-z||_2^2}{2\sigma^2} = \exp -\gamma||x-z||_2^2$$

Noyau Laplacien

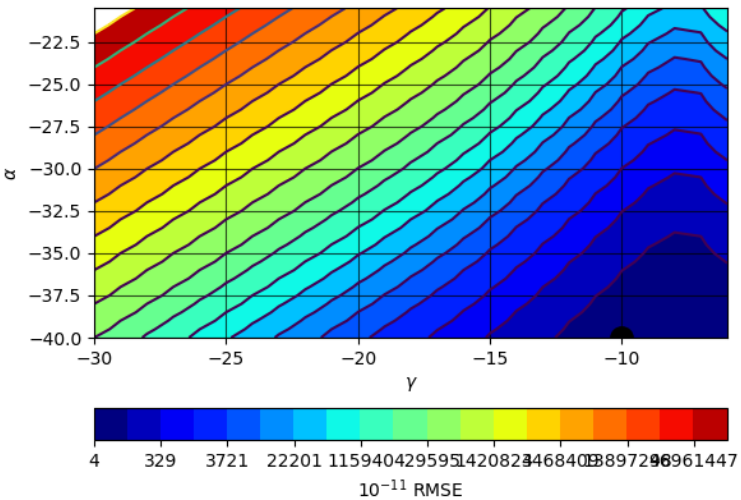
$$K(x,z) = \exp -\frac{||x-z||_1}{\sigma} = \exp -\gamma||x-z||_1$$

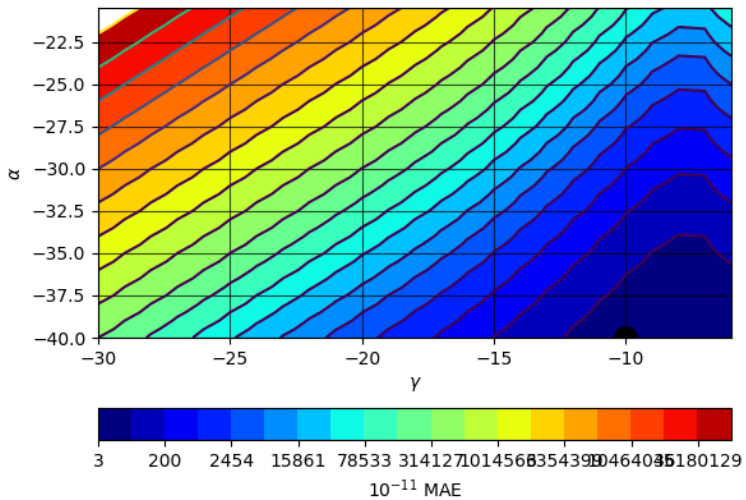
Résultats noyau gaussien





Résultats noyau laplacien





Recherche hyparamètres optimaux

Erreur utilisée	RMSE	MAE	R^2
KRR avec noyau gaussien	9.6371	7.6962	0.9977
KRR avec noyau laplacien	5.4019	3.4555	0.9993

Perspectives

- Utilisation d'autres modèle comme Random Forest
- Utilisation d'autres descripteurs que les matrices de Coulomb

Conclusion

- Un projet efficient pour développer ces compétences en machine learning.

- Nous avons reproduis l'article scientifique
- Nous ne sommes pas allé plus loin...

Merci de votre attention