© Théo Jalabert

Introduction à l'Apprentissage Statistique
-
SVM

M2 Actuariat – ISFA – 2021/2022

Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

In God we trust, all others bring data

- William Edwards Deming

© Théo Jalabert

Noyaux

© Théo Jalabert

Régression

Régression entre Y et X

$$\mathbb{E}[Y \mid X = x] = m(x) + \epsilon$$

Dans le cas de la régression linéaire on impose une forme affine

$$m(x) = a \times x + b$$

puis on cherche les paramètres a et b

Dans la régression non paramétrique on étend cette approche en exprimant la fonction m à partir de la base d'entrainement

$$m(x) = \sum_{i=1}^{n} w(x_i, x) \times y_i$$

© Théo Jalabert

Régression par noyaux

Noyau est une fonction réelle

- Positive
- Symétrique
- Intégrable

$$\int_{-\infty}^{+\infty} K(u)du = 1 \quad \text{et} \quad K(-u) = K(u)$$

On définit alors les poids de la régression non paramétrique

$$w_h(x_i, x) = \frac{K_h(x_i - x)}{\sum_i^n K_h(x_i - x)}$$

© Théo Jalabert

Noyaux usuels

Uniforme

$$\frac{1}{2}\mathbb{1}\left(\frac{|u|}{h} \leq 1\right)$$

LOESS

$$\left(1 - \frac{|u|^3}{h^3}\right) \times \mathbb{1}\left(\frac{|u|}{h} \leq 1\right)$$

Triangle

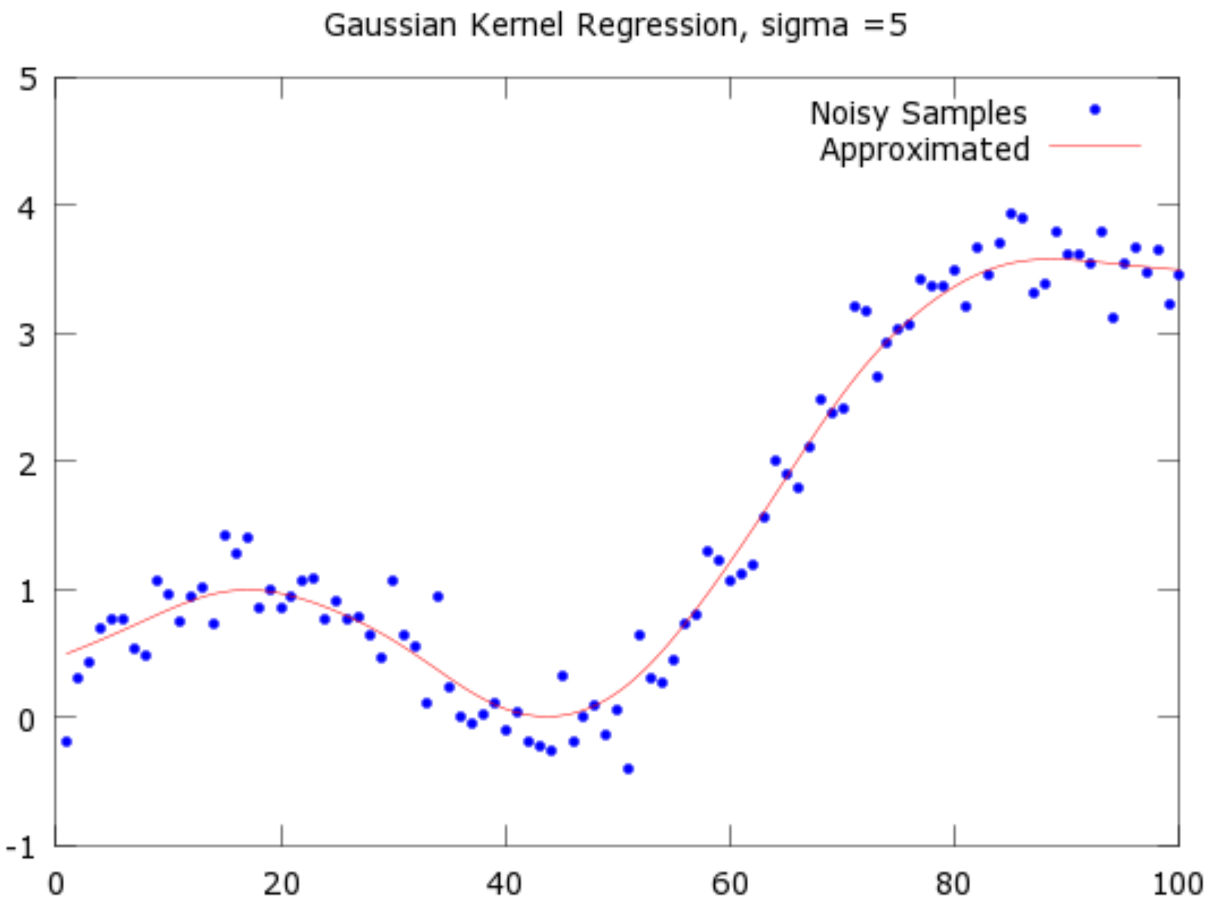
$$\left(1 - \frac{|u|}{h}\right) \times \mathbb{1}\left(\frac{|u|}{h} \leq 1\right)$$

Gaussien

$$\frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{1}{2}u^2 h^{-2}\right)$$

© Théo Jalabert

Exemple



© Théo Jalabert

SVM Linéaire

© Théo Jalabert

Support Vector Machine

Soit un échantillon d'entrainement $\{\boldsymbol{x}_i, y_i\}_{i=1,\dots,n}$ avec $\boldsymbol{x} \in \mathbb{R}^p$ et $y \in \{-1, +1\}$

- on cherche à créer un modèle de classification de y

Principe du SVM : trouver un hyperplan dans $\mathbb{R}^p$ qui sépare les données

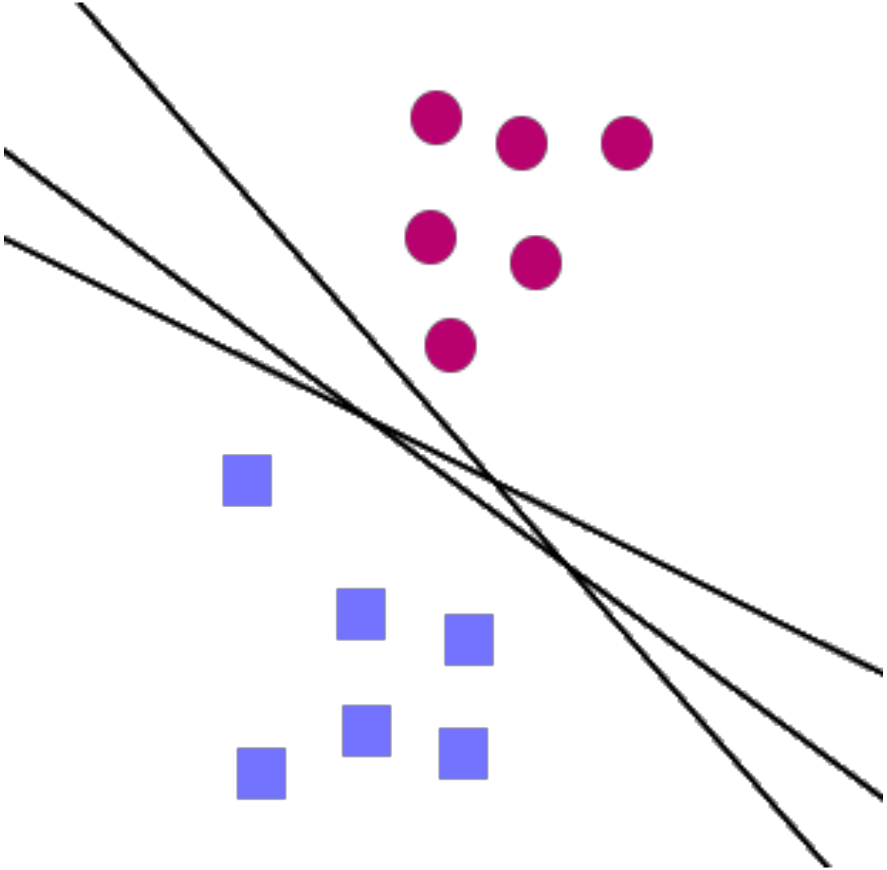
$$< \boldsymbol{\omega}, \boldsymbol{x} > + b = \boldsymbol{\omega}^t \boldsymbol{x} + b = 0$$

Pour classifier un nouveau point x_0, le SVM estimera alors

$$\text{signe}(\boldsymbol{\omega}^t \boldsymbol{x}_0 + b)$$

© Théo Jalabert

Support Vector Machine



© Théo Jalabert

Maximisation de la marge

Plusieurs hyperplans peuvent correspondre

- Quel hyperplan choisir ?
- Quel programme d'optimisation retenir ?

L'hyperplan qui maximise la distance au point le plus proche de chaque classe : on parlera de marge maximale

$$\min_{\boldsymbol{\omega},b} \boldsymbol{\omega}^t \boldsymbol{\omega}$$

sous contraintes

$$\forall i, y_i(\boldsymbol{\omega}^t \boldsymbol{x}_i + b) \geq 1$$

© Théo Jalabert

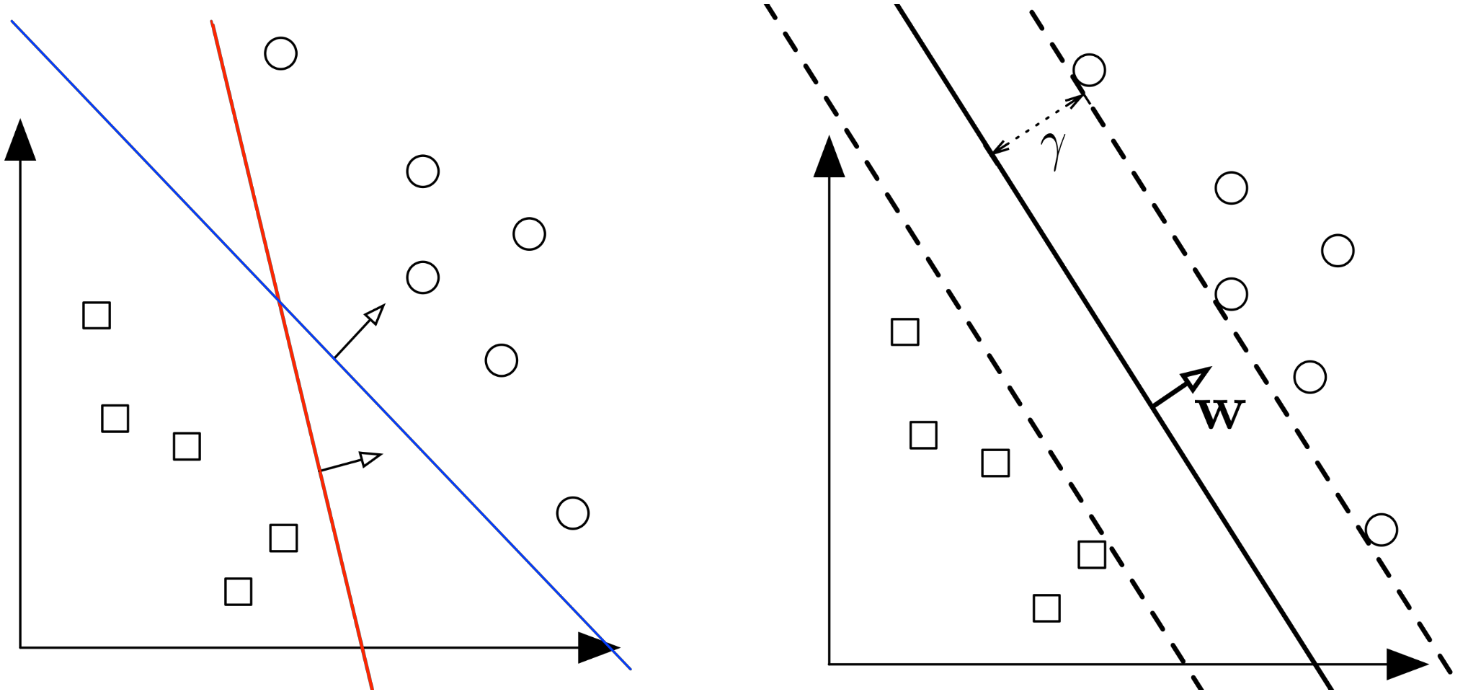
Preuve

Trois étapes

1. Définir la marge en fonction de $\boldsymbol{\omega}$, b et $\boldsymbol{x}$
2. Maximisation de la marge pour l'hyperplan séparateur
3. Simplification du programme d'optimisation

Marge maximale

© Théo Jalabert



© Théo Jalabert

SVM à contraintes souples

Problème : dans la vraie vie les groupes ne sont généralement pas séparables

- Autoriser les mauvais classements…
- … pénalisés par un coût C

Définition des variables ressorts ξ_i correspondant aux erreurs de classification

$$\min_{\boldsymbol{\omega},b} \boldsymbol{\omega}^t\boldsymbol{\omega} + C\sum_{i}^{n} \xi_i$$

sous contraintes

$$\forall i, y_i(\boldsymbol{\omega}^t\boldsymbol{x}_i + b) \geq 1 - \xi_i$$
$$\forall i, \xi_i \geq 0$$

© Théo Jalabert

Souplesse de la marge

Le coût C de mauvais classement est un hyperparamètre à tuner !

Si le coût C est élevé

- Modèle très stricte et recherche d'une solution complexe
- Meilleur ajustement mais risque d'overfitting

Si le coût C est faible

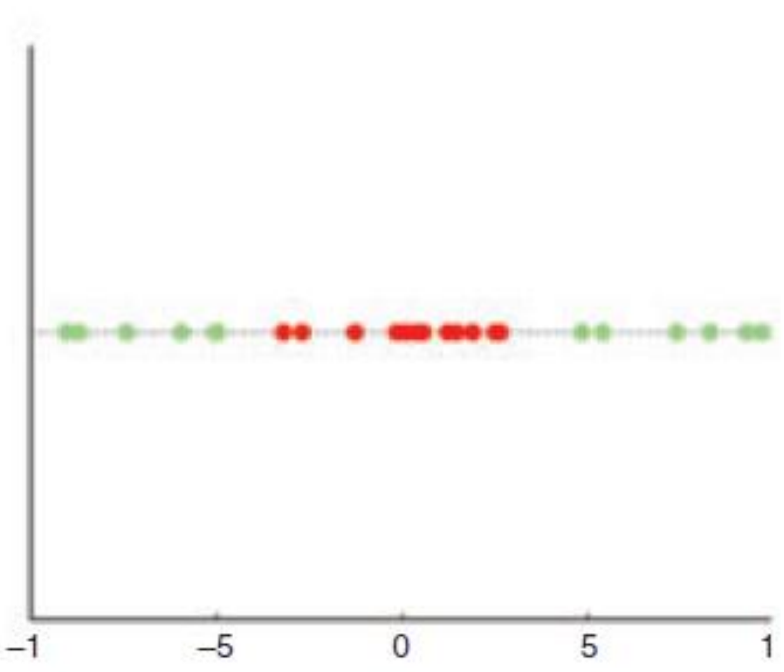
- Beaucoup d'erreurs tolérées sur la base d'entrainement
- On sacrifie certains points pour avoir une solution plus simple

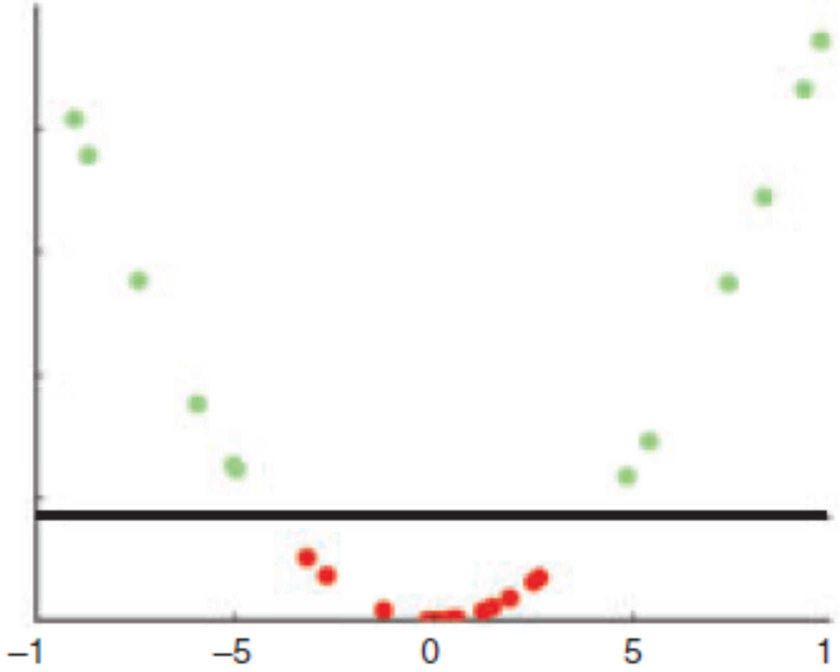
© Théo Jalabert

SVM à Noyaux

© Théo Jalabert

Changement de dimension





© Théo Jalabert

Principe du SVM à noyaux

Jusqu'à présent : problèmes linéairement séparables (ou presque)

Solution

- Plonger les données dans un nouvel espace de dimension supérieure
- Puis appliquer un SVM linéaire dans ce nouvel espace

L'astuce est d'utiliser une fonction noyaux

- Pas besoin de connaître explicitement le nouvel espace
- Juste besoin de connaître le noyaux pour le problème d'optimisation

© Théo Jalabert

Noyaux

Le noyaux est une fonction positive semi-définie

$$K : E \times E \ \rightarrow \ \mathbb{R}$$

telle que

$$K(x, x') = \phi(x)^t \phi(x')$$

où $x \ \rightarrow \phi(x)$ est la projection dans l'espace de dimension supérieur

Le problème d'optimisation devient alors

$$\min_{\boldsymbol{\omega}, b} \boldsymbol{\omega}^t \boldsymbol{\omega} + C \sum_{i}^{n} \xi_i$$

sous contraintes

$$\forall i, y_i(\boldsymbol{\omega}^t \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i$$

$$\forall i, \xi_i \geq 0$$

© Théo Jalabert

Noyaux usuels pour SVM

Linéaire

$$K(x, x') = x^t x'$$

Radial Basis Function (RBF)

$$K(x, x') = \exp\left(-\frac{(x - x')^2}{\gamma^2}\right)$$

Polynomial

$$K(x, x') = (\gamma x^t x' + c_0)^d$$

Sigmoid

$$K(x, x') = \tanh(\gamma x^t x' + c_0)$$

© Théo Jalabert

En pratique

© Théo Jalabert

Data pre-processing

Scaling

- Normaliser les variables numériques permet d'avoir de meilleures performances
- Eviter que certaines variables avec de grandes valeurs ne prennent trop d'importance
- L'utilisation du noyaux a tendance augmente les problématiques de calculs

Encodage binaire (*one hot*) des variables catégorielles

- Le SVM a besoin de variables numériques car algorithme purement géométrique
- Pour les variables catégorielles à m valeurs, il faut créer m variables binaires

© Théo Jalabert

Choix du modèle

Choix du noyaux

- RBF est un bon choix par défaut
- Non linéaire, bonne performance en cas général, seulement 2 hyperparamètres

Tuning

- Deux hyperparamètres : γ et C
- Grid search + cross-validation

Attention au besoin en puissance de calcul

- Même si l'algorithme semble simple, il est très gourmand en calcul
- $O(N^3)$ en temps et $O(N^2)$ en espace

© Théo Jalabert

Extensions

© Théo Jalabert

Multiclasses

On a étudié le SVM pour un classification binaire, mais il est possible de l'étendre aux problématiques à m classes

Un contre tous

- Entrainer un SVM binaire pour chaque problème $\{i \in classe_j, i \notin classe_j\}$
- Résolution de m SVM différents

Un contre un

- Entrainer un SVM binaire pour chaque problème $\{i \in classe_j, i \in classe_k\}$
- Résolution de $\frac{m(\mathrm{m}-1)}{2}$ SVM différents

© Théo Jalabert

Régression : SVR

On se place maintenant dans le cas où $y_i \in \mathbb{R}$

- on ne cherche pas un hyperplan qui sépare au mieux les données
- mais un hyperplan qui approche au mieux les valeurs y_i

On souhaite que les y_i soient dans une bande de largeur 2ϵ autour de l'hyperplan, avec $\epsilon > 0$ petit, i.e.

$$|\boldsymbol{\omega}^t \boldsymbol{x}_i + b - y_i| < \epsilon$$

© Théo Jalabert

SVR - Optimisation

Par analogie avec le SVM, le problème d'optimisation du SVR linéaire est

$$\min_{\boldsymbol{\omega},b} \boldsymbol{\omega}^t \boldsymbol{\omega}$$

sous contraintes

$$\forall i, |y_i - \boldsymbol{\omega}^t \boldsymbol{x}_i - b| \leq \epsilon$$

De même on rajoute de la souplesse à la marge avec des variables ressorts et on projette le problème dans un espace de dimension supérieur avec un noyaux

$$\min_{\boldsymbol{\omega},b} \boldsymbol{\omega}^t \boldsymbol{\omega} + C + \sum_{i=1}^{n} |\xi_i|$$

sous contraintes

$$\forall i, |y_i - \boldsymbol{\omega}^t \phi(\boldsymbol{x}_i) - b| \leq \epsilon + |\xi_i|$$