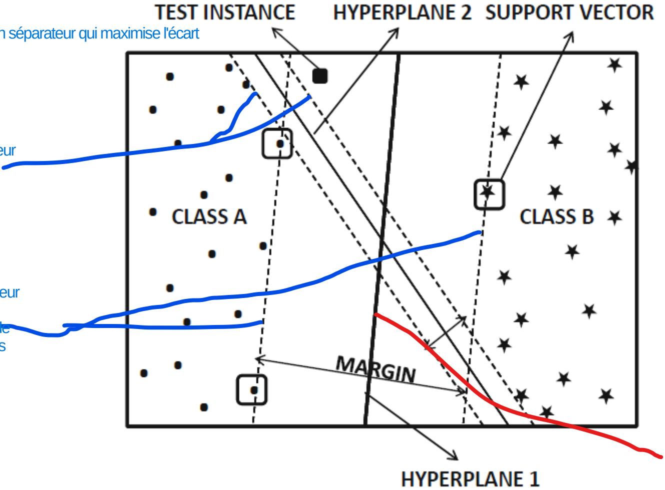
Machines à vecteurs de support

Vecteur de support de machine: SVM

- Les SVMs travaillent en recherchant des séparateurs linéaires de marge maximum entre deux classes on veut un séparateur qui maximise l'écart
- Lorsque les données sont linéairement séparables, il existe un nombre infini de façons possibles de construire un hyperplan séparateur linéaire entre les classes

on veut un séparateur qui maximise l'écart

séparateur



séparateur + large, moins de chances de faire erreur. Plus robuste

plan séparateur

Aggarwal, 2015

- Lequel hyperplan est le meilleur ?

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Machines à vecteurs de support

- La raison de la variation des performances des deux classifieurs est que l'instance de test est placée dans une région limite entre les deux classes
- Dans une telle situation, le classifieur avec le plus grande marge est préféré $\Rightarrow$ plus robuste
- Les points de données d'entraînement sur les hyperplans parallèles à l'hyperplan séparateur classifieur sont appelés les vecteurs de support
- L'hyperplan séparateur est précisément au milieu de ces deux hyperplans afin d'atteindre la classification la plus précise

Machines à vecteurs de support

SAVOIR LAGRANGE

- Comment détermine-t-on l'hyperplan de marge maximale ?
- L'hyperplan séparateur est donné par :

$$w^T X + b = 0$$

- Les $d+1$ coefficients de w et b doivent être appris à partir des données d'apprentissage pour maximiser la marge de séparation entre les deux classes

Machines à vecteurs de support

- On assume que :
  1. les données d'entraînement sont linéairement séparables
  2. les classes sont étiquetées +1 et -1
- Alors,

équation de l'hyperplan

$$w^T X_i + b \geq 0 \quad \forall i : y_i = +1$$

$$w^T X_i + b \leq 0 \quad \forall i : y_i = -1$$

- Ces contraintes ne définissent pas la marge

Machines à vecteurs de support

- Le problème d'optimisation qu'on veut maximiser peut être exprimé comme :

maximiser la M marge

M: distance

$$\max_{w,b,\|w\|=1} \quad M$$
$$s.t. \text{ tel que} \quad y_i(w^T X_i + b) \geq M \quad \forall i$$

- Nous pouvons enlever la contrainte $\|w\| = 1$ en faisant

$$\frac{1}{\|w\|} y_i(w^T X_i + b) \geq M$$

Machines à vecteurs de support

- Puisque pour n'importe quels w et b qui satisfont les inégalités, ses multiples positives les satisfont eux aussi, nous pouvons fixer $\|w\| = 1/M$
- Ce qui implique :

$$y_i(w^T X_i + b) \geq 1 \quad \forall i$$

- On peut montrer avec l'algèbre linéaire que la distance M entre les deux hyperplans parallèles est égale à $2/\|w\|$
- Maximiser cette distance est donc équivalent à minimiser $\|w\|^2/2$ (problème d'optimisation quadratique – plus simple).

Machines à vecteurs de support

SVM linéaire

$$\begin{aligned} \min \quad & \|w\|^2/2 \\ s.t. \quad & y_i(w^T X_i + b) \geq 1 \quad \forall i \end{aligned}$$

Remarque que chaque donnée d'entraînement ajoute une contrainte au modèle...

ex: individu 1
individu 2

Machines à vecteurs de support

- De tels problèmes d'optimisation non linéaires avec des contraintes sont résolus en utilisant une méthode connue sous le nom de relaxation Lagrangienne
- L'idée générale est d'associer un ensemble n de multiplicateurs non-négatifs $\lambda = (\lambda_1 \dots \lambda_n) \geq 0$ pour les différentes inégalités de la marge
- Les contraintes sont ensuite relaxées et la fonction objectif est augmentée en incorporant une pénalité lagrangienne

$$L_P = \|w\|^2/2 - \sum_{i=1}^{n} \underbrace{\lambda_i}_{\text{pénaliser l'erreur}} [\underbrace{y_i(w^T X_i + b) - 1}_{\text{contraintes}}]$$

Machines à vecteurs de support

- Si L_P est minimisé par rapport à w et b pour un λ particulier,
trouver les plus gros coefficients
puis maximisée par rapport aux multiplicateurs λ, la solution duale résultante L_D est une borne inférieure de la fonction objectif optimale O de la SVM linéaire

$$O \geq L_D = \max_{\lambda \geq 0} \min_{w,b} L_P$$

SVM vs regression logistique à l'Examen!

Savoir la différence

Machines à vecteurs de support

- Si L_P est minimisé par rapport à w et b pour un λ particulier, puis maximisée par rapport aux multiplicateurs λ, la solution duale résultante L_D est une borne inférieure de la fonction objectif optimale O de la SVM linéaire

$$O \geq L_D = \max_{\lambda \geq 0} \min_{w,b} L_P$$

- Dans le cas de la SVM linéaire, ce résultat est encore plus fort puisque la f.o. est convexe
- c.a.d $O = L_D$

Machines à vecteurs de support

- En utilisant les conditions d'optimalité de premier ordre, c'est possible prouver que

EXAMEN
formules pour expliquer les choses: KErnell Trick

objectif

$$O = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j X_i \cdot X_j$$

s.t. $\lambda \geq 0$ and $\sum_{i=1}^{n} \lambda_i y_i = 0$

!

- Ainsi que

paramètres w qu'on essaye de trouver

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

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Machines à vecteurs de support

- Pour une instance de test Z, sa classe $F(Z)$ est défini par

$$F(Z) = sign\{w^T Z + b\} = sign\{(\sum_{i=1}^{n} \lambda_i y_i X_i \cdot Z) + b\}$$

- Remarque que la classification ainsi que la valeur de la f.o. de la SVM sont obtenus sans qu'on ait besoin de connaître la matrice de données X directement
- Tout est exprimé par des produits scalaires
- Cela est crucial pour l'application de la SVM pour des données pas linéairement séparables (kernel trick)

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

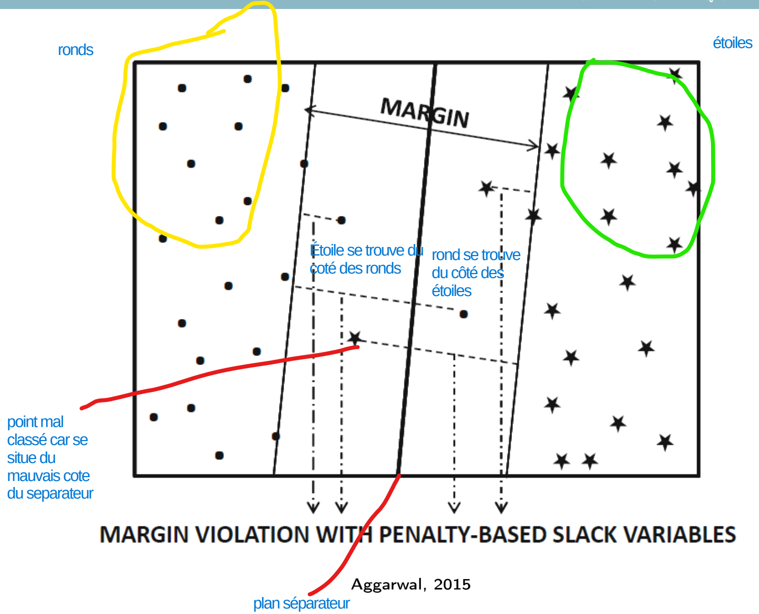
Machines à vecteurs de support

EXAMEN

ajouter un terme de pénalité

SVM linéaire avec marge légère pour données non lin.sep.

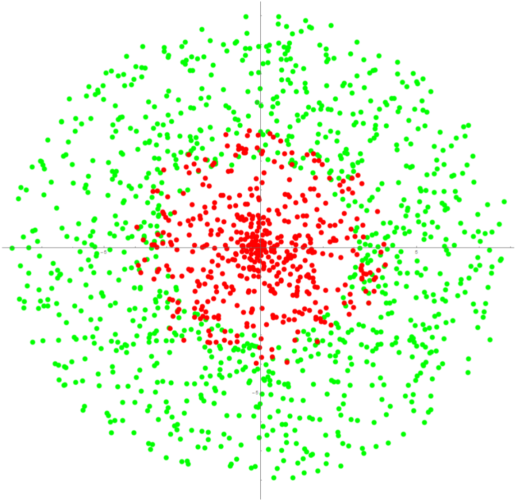
$$
\begin{array}{lll}
\min & \|w\|^2/2 + C\sum_{i=1}^{n}\xi_i & \\
s.t. & WX_i + b \geq +1 - \xi_i & \forall i : y_i = +1 \\
 & WX_i + b \leq -1 + \xi_i & \forall i : y_i = -1 \\
 & \xi_i \geq 0 & \forall i
\end{array}
$$



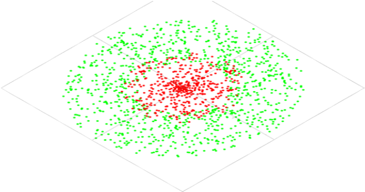
Aggarwal, 2015

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

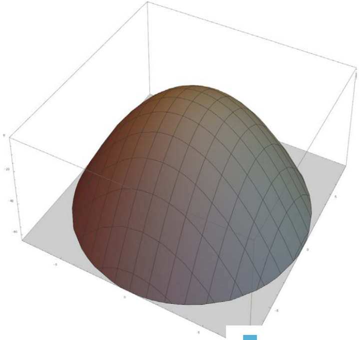
Projections



Projections

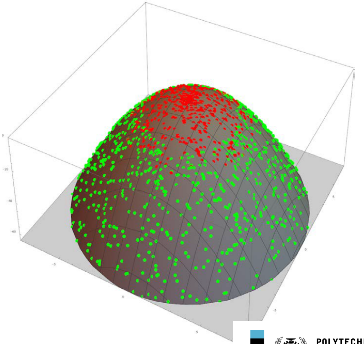


$$(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$$



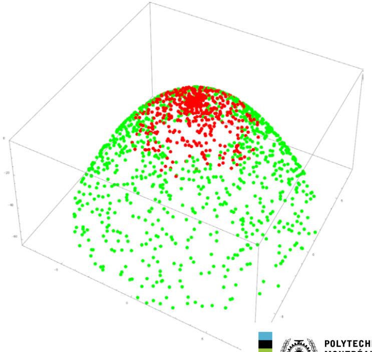
Projections

$$(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$$



Projections

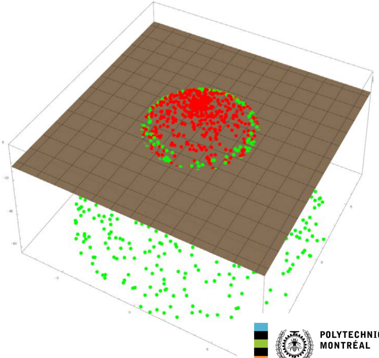
$$(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$$



POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Projections

$$(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$$

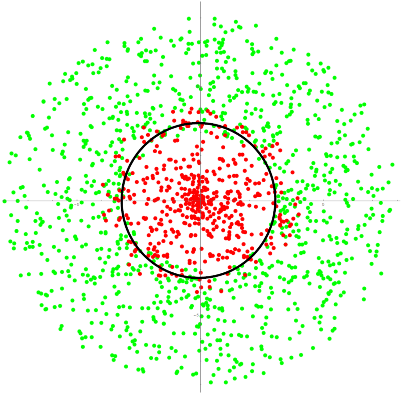


POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Projections

- La non-linéarité du classifieur dépend de comment l'espace est projeté à des dimensions plus élevées

- Le "truc" des SVMs est qu'il n'est pas nécessaire de connaître les features des données dans le nouveau espace



SVMs non-linéaires

- la formulation SVM peut être entièrement résolue en termes de produits scalaires (ou similarités) entre des paires de points de données
- la clé est de définir le produit scalaire (ou la fonction de similarité) directement dans la représentation transformée $\Phi(X)$ avec l'utilisation d'une fonction noyau $\mathcal{K}(X_i, X_j)$

$$\mathcal{K}(X_i, X_j) = \Phi(X_i) \cdot \Phi(X_j)$$

SVMs non-linéaires

- Cela implique :

$$O = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \mathcal{K}(X_i, X_j)$$

et

$$F(Z) = sign\{WZ + b\} = sign\{(\sum_{i=1}^{n} \lambda_i y_i \mathcal{K}(X_i, Z)) + b\}$$

SVMs non-linéaires

- Tous les calculs sont effectués dans l'espace d'origine avec la fonction noyau choisie
- La transformation réelle Φ n'a pas besoin d'être connue tant que la fonction noyau $\mathcal{K}$ soit connue
- En utilisant la similarité basée sur des noyaux choisis avec soin, on obtient des SVMs non-linéaires

SVMs non-linéaires

- Il y a différentes manières de modéliser la similarité entre les données avec des noyaux
- Chacune a des avantages sur certains ensembles de données, il est donc nécessaire de jouer avec les options disponibles dans les librairies pour obtenir les meilleures performances
- ex.

Function	Form
Gaussian radial basis kernel	$K(\overline{X_i}, \overline{X_j}) = e^{-\|\overline{X_i}-\overline{X_j}\|^2/2\sigma^2}$
Polynomial kernel	$K(\overline{X_i}, \overline{X_j}) = (\overline{X_i} \cdot \overline{X_j} + c)^h$
Sigmoid kernel	$K(\overline{X_i}, \overline{X_j}) = \tanh(\kappa \overline{X_i} \cdot \overline{X_j} - \delta)$

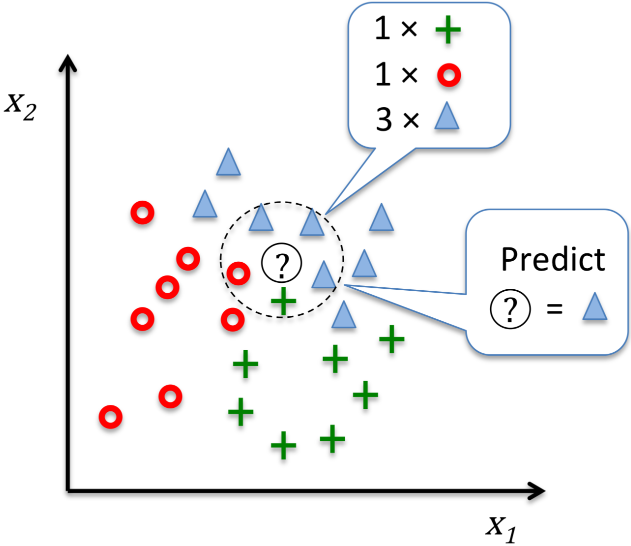
PAs besoin de savoir ca, mais expliquer cmt cela fonctionne à l'examen!!

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

k plus proches voisins

1. Trouver les k enregistrements de X qui sont les plus proches de la nouvelle donnée Z que l'on veut classifier
2. Faire voter chacune de ces enregistrements selon leurs classes associées y
3. Retourner la classe majoritaire

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE



source : Raschka, 2015

k plus proches voisins

- Le succès de l'algorithme va dépendre de deux facteurs principaux :
  - la quantité de données d'entraînement
  - la qualité de la mesure de distance (deux enregistrements similaires doivent faire partie de la même classe)

- Grand pouvoir de représentation
  - ⋆ peut classifier des données non linéairement séparables

- Par contre, performance liée à notre capacité de stockage