EE-311—Apprentissage et intelligence artificielle

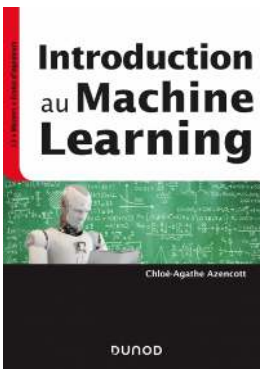
4. Machines à vecteurs de support

Michael Liebling

https://moodle.epfl.ch/course/view.php?id=16090

14 mars 2025 (compilé le 13 mars 2025)

Ouvrage de référence et source



Ces transparents sont basés en grande partie sur le texte de Chloé-Agathe Azencott "Introduction au Machine Learning", Dunod, 2019
ISBN 978-210-080153-4

L'auteure a mis le texte (sans les exercices) à disposition ici : http://cazencott.info/dotclear/public/lectures/IntroML_Azencott.pdf

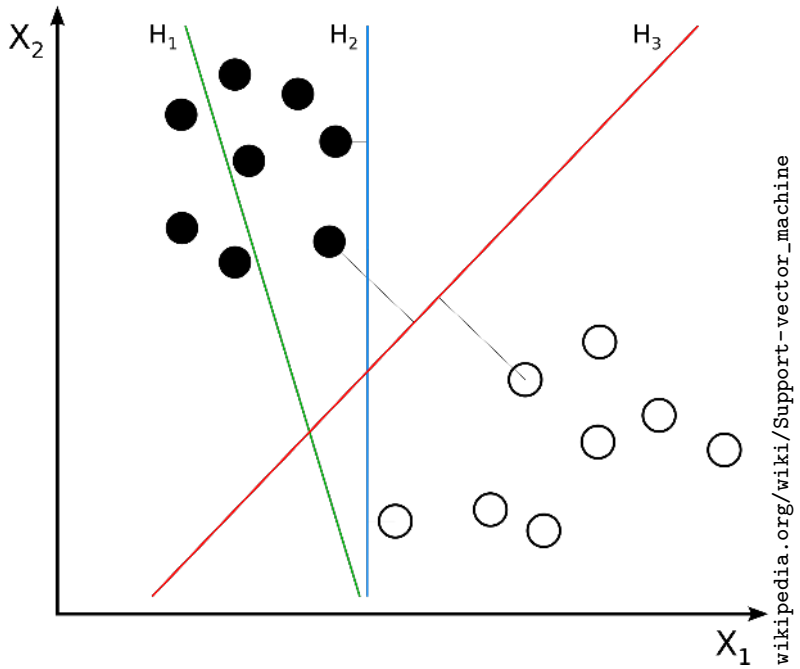
Avertissement : Bien que ces transparents partagent la notation mathématique, la structure de l'exposition (en partie), et certains exemples avec le livre, ils ne constituent qu'un complément et non un remplacement ou une source unique pour la couverture des matières du cours. À ce titre, ces transparents ne se substituent pas au texte.

Contenu

- Préliminaires : historique, séparabilité linéaire, hyperplans, marges
- Machines à vecteurs de support à marge rigide
- Machines à vecteurs de support à marge souple
- Cas non linéaire : SVM à noyau

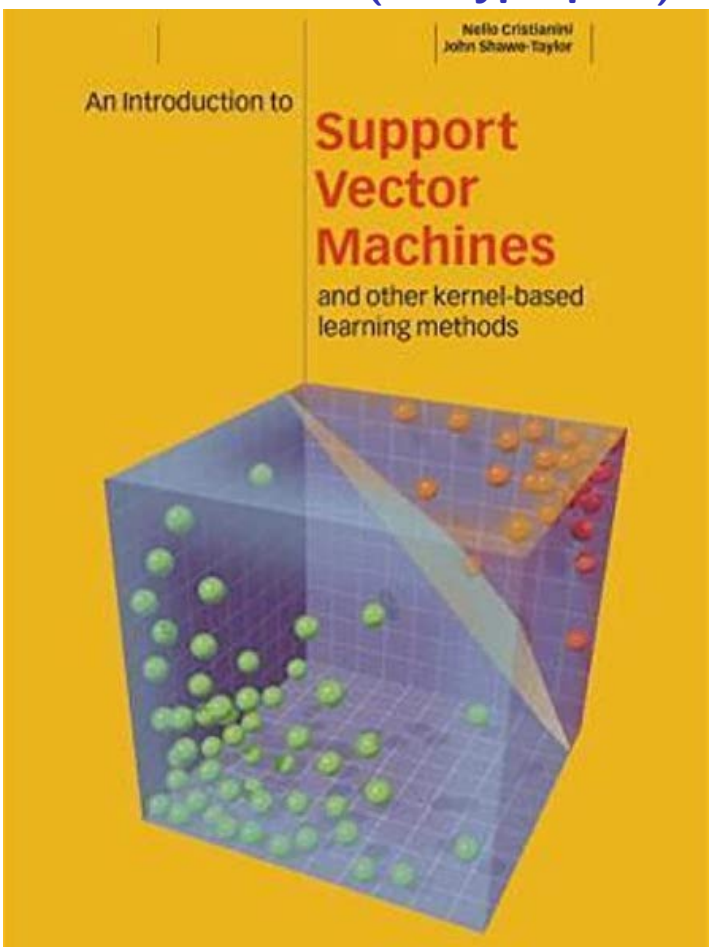
Support vector machines (SVM) = Machines à vecteurs de support

Motivation : trouver un hyperplan qui sépare 2 classes



H_1 : ne sépare pas les deux classes
H_2 : séparation mais faible marge
H_3 : séparation avec marge maximale ← obtenu via algorithme SVM

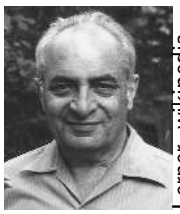
Séparation en dimensions > 2 ($\Rightarrow$ hyperplan)



Historique des machines à vecteurs de support

Les machines à vecteurs de support (aussi appelées machines à vecteurs supports), ou SVM de l'anglais *support vector machines* se basent sur un algorithme linéaire proposé par Vladimir Vapnik et Aleksandr Lerner en 1963 (Vapnik et Lerner, 1963)


Vapnik, Columbia U.


Lerner, wikipedia

Étendues efficacement à l'apprentissage de modèles non linéaires grâce à l'astuce du noyau par Vladimir Vapnik, Bernhard Boser, Isabelle Guyon et Corinna Cortes (Boser et al., 1992 ; Cortes et Vapnik, 1995)


Boser, UC Berkeley

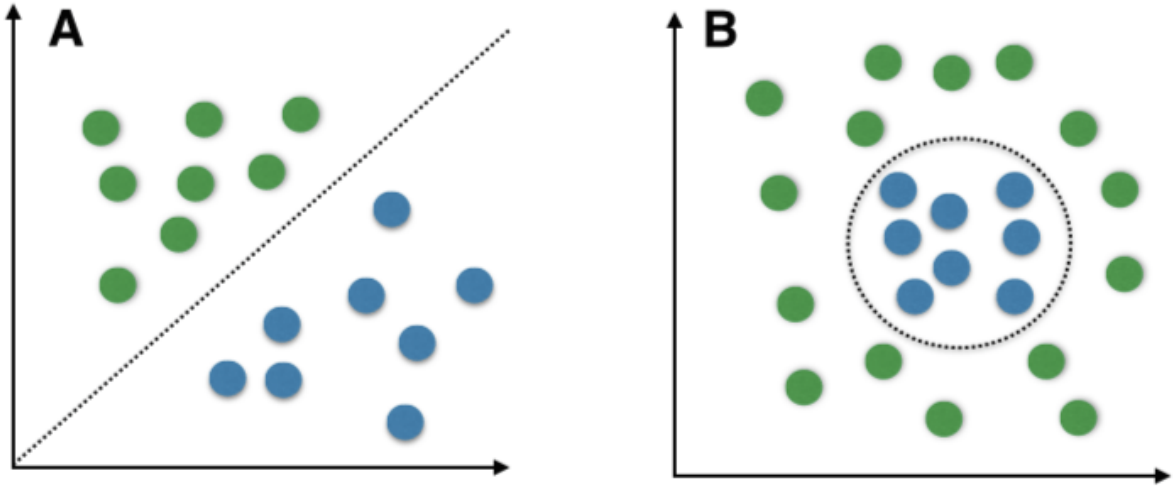

Guyon, homepage


Cortes, Google

Définition 10.1 (Séparabilité linéaire)

Définition 10.1 (Séparabilité linéaire) Soit $\mathscr{D} = \left\{\vec{x}^i, y^i\right\}_{i=1,\ldots,n}$ un jeu de données de n observations. Nous supposons que $\vec{x}^i \in \mathbb{R}^p$ et $y^i \in \{-1, 1\}$. On dit que $\mathscr{D}$ est linéairement séparable s'il existe au moins un hyperplan dans $\mathbb{R}^p$ tel que tous les points positifs (étiquetés $+1$) soient d'un côté de cet hyperplan et tous les points négatifs (étiquetés -1) de l'autre.

Linéairement séparable **Pas séparable (linéairement)**

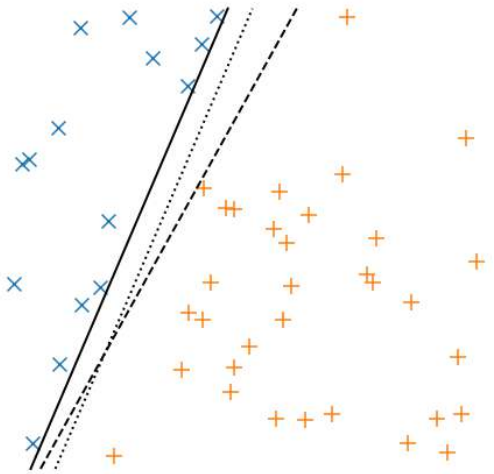


http://sebastianraschka.com/Articles/2014_naive_bayes_1.html

Illustration du problème de classification

Étant donné un jeu d'entraînement linéairement séparable :

- Il peut exister une infinité d'hyperplans séparateurs qui ne font aucune erreur de classification
- Ces hyperplans sont des modèles équivalents du point de vue de la minimisation du risque empirique

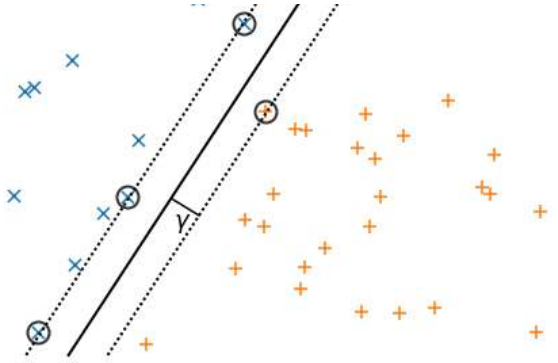


Azencott

FIGURE 10.1 – Une infinité d'hyperplans (en deux dimensions, des droites) séparent les points négatifs (x) des points positifs (+).

Marge d'un hyperplan séparateur (données lin. séparable)

Définition 10.2 (Marge) La marge γ d'un hyperplan séparateur est la distance de cet hyperplan à l'observation du jeu d'entraînement la plus proche.



Azencott

Nous cherchons donc l'hyperplan qui maximise la marge.

Note : Il y a au moins une observation négative et une observation positive à une distance γ de l'hyperplan séparateur (si par exemple toutes les observations négatives étaient à une distance supérieure à γ de l'hyperplan séparateur, on pourrait rapprocher cet hyperplan des observations négatives et augmenter la marge).

Vecteurs de support

Définition 10.3 (Vecteurs de support) On appelle vecteurs de support les observations du jeu d'entraînement situés à une distance γ de l'hyperplan séparateur. Elles «soutiennent» les hyperplans H_+ et H_-.

Note : Toutes les observations positives sont situées à l'extérieur de H_+, tandis que toutes les observations négatives sont situées à l'extérieur de H_-.

Origine du nom SVM :

- Support Vector Machine (SVM, machine à vecteurs de support)
- séparatrice à vaste marge

Zone d'indécision

Définition 10.4 (Zone d'indécision) On appelle zone d'indécision la zone située entre H_- et H_+. Cette zone ne contient aucune observation.

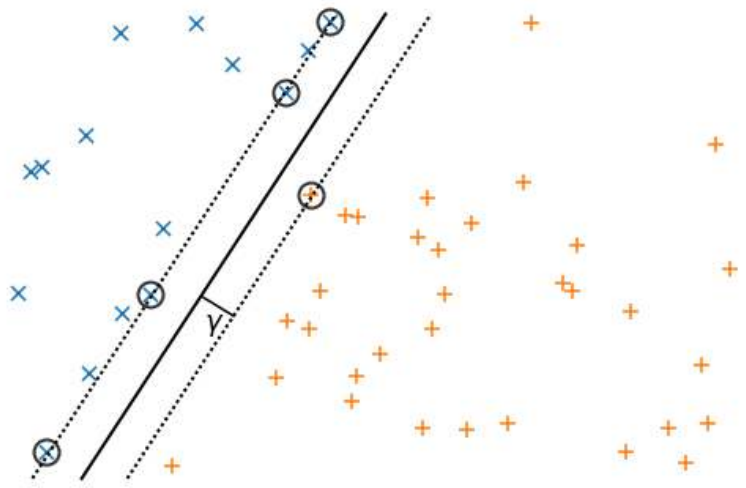


FIGURE 10.2 – La marge γ d'un hyperplan séparateur (ici en trait plein) est sa distance à l'observation la plus proche. Quand cette marge est maximale, au moins une observation négative et une observation positive sont à une distance γ de l'hyperplan séparateur. Les hyperplans (ici en pointillés) parallèles à l'hyperplan séparateur et passant par ces observations définissent la zone d'indécision. Les observations situées sur ces hyperplans (cerclées) sont les vecteurs de support. Azencott

Formulation de la SVM à marge rigide

L'équation de l'hyperplan séparateur H que nous cherchons est de la forme

$$\langle \vec{w}, \vec{x} \rangle + b = 0,$$

où $\langle \, , \rangle$ représente le produit scalaire sur $\mathbb{R}^p$

L'hyperplan H_+ est parallèle à H et de la forme :

$$\langle \vec{w}, \vec{x} \rangle = \text{constante}$$

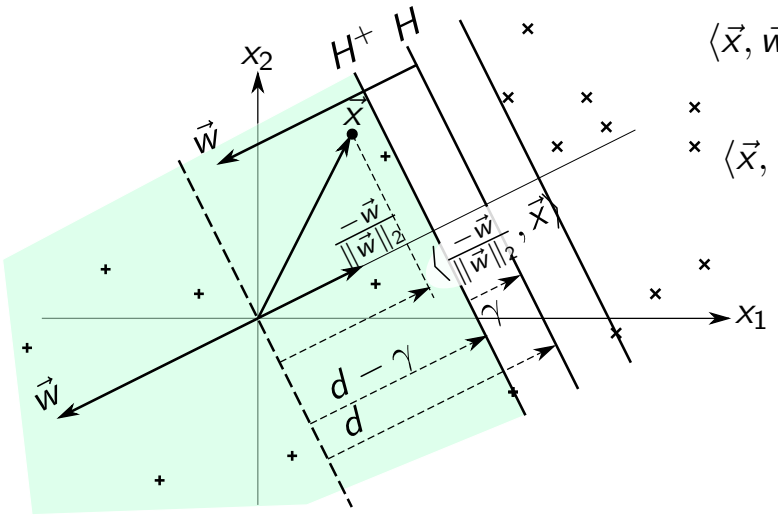
Nous pouvons fixer cette constante à 1 (sans perte de généralité, il suffirait d'ajuster $\vec{w}$ et b proportionnellement pour d'autres choix) et on a :

$$H_+ : \langle \vec{w}, \vec{x} \rangle + b = 1$$
$$H_- : \langle \vec{w}, \vec{x} \rangle + b = -1$$
$$\text{Marge} : \gamma = \frac{1}{\|\vec{w}\|_2}$$

Dérivation : observations positives

Pour les observations positives

$$\left\langle \vec{x}, \frac{-\vec{w}}{\|\vec{w}\|_2} \right\rangle \leq d - \gamma \quad \Big| \cdot (-\|\vec{w}\|_2)$$



$$\langle \vec{x}, \vec{w} \rangle \geq -d\|\vec{w}\|_2 + \gamma\|\vec{w}\|_2$$

$$\langle \vec{x}, \vec{w} \rangle + \underbrace{d\|\vec{w}\|_2}_{b} \geq \underbrace{\gamma\|\vec{w}\|_2}_{1}$$

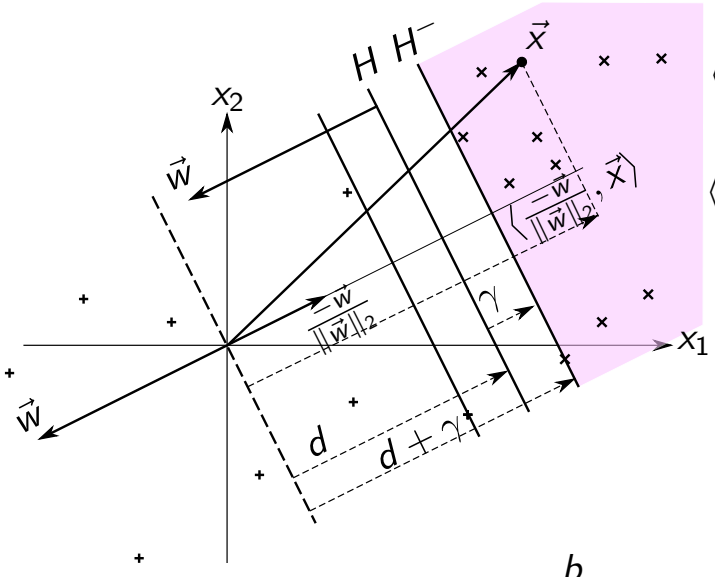
$$\langle \vec{x}, \vec{w} \rangle + b \geq 1$$

$$d = \frac{b}{\|\vec{w}\|_2} \quad \gamma = \frac{1}{\|\vec{w}\|_2}$$

Dérivation : observations négatives

Pour les observations négatives :

$$\left\langle \vec{x}, \frac{-\vec{w}}{\|\vec{w}\|_2} \right\rangle \geq d + \gamma \quad \Big| \cdot (-\|\vec{w}\|_2)$$



$$\langle \vec{x}, \vec{w} \rangle \leq -d\|\vec{w}\|_2 - \gamma\|\vec{w}\|_2$$

$$\langle \vec{x}, \vec{w} \rangle + \underbrace{d\|\vec{w}\|_2}_{b} \leq -\underbrace{\gamma\|\vec{w}\|_2}_{1}$$

$$\langle \vec{x}, \vec{w} \rangle + b \leq -1$$

$$d = \frac{b}{\|\vec{w}\|_2} \quad \gamma = \frac{1}{\|\vec{w}\|_2}$$

Dérivation des équations et grandeurs (résumé)

Pour les points négatifs :

$$\left\langle \vec{x}, \frac{-\vec{w}}{\|\vec{w}\|_2} \right\rangle \geq d+\gamma \quad \Big| \cdot(-\|\vec{w}\|_2)$$

$$\langle \vec{x}, \vec{w} \rangle \leq -d\|\vec{w}\|_2 - \gamma\|\vec{w}\|_2$$

$$\langle \vec{x}, \vec{w} \rangle + \underbrace{d\|\vec{w}\|_2}_{b} \leq -\underbrace{\gamma\|\vec{w}\|_2}_{1}$$

$$\langle \vec{x}, \vec{w} \rangle + b \leq -1$$

Pour les points positifs

$$\left\langle \vec{x}, \frac{-\vec{w}}{\|\vec{w}\|_2} \right\rangle \leq d-\gamma \quad \Big| \cdot(-\|\vec{w}\|_2)$$

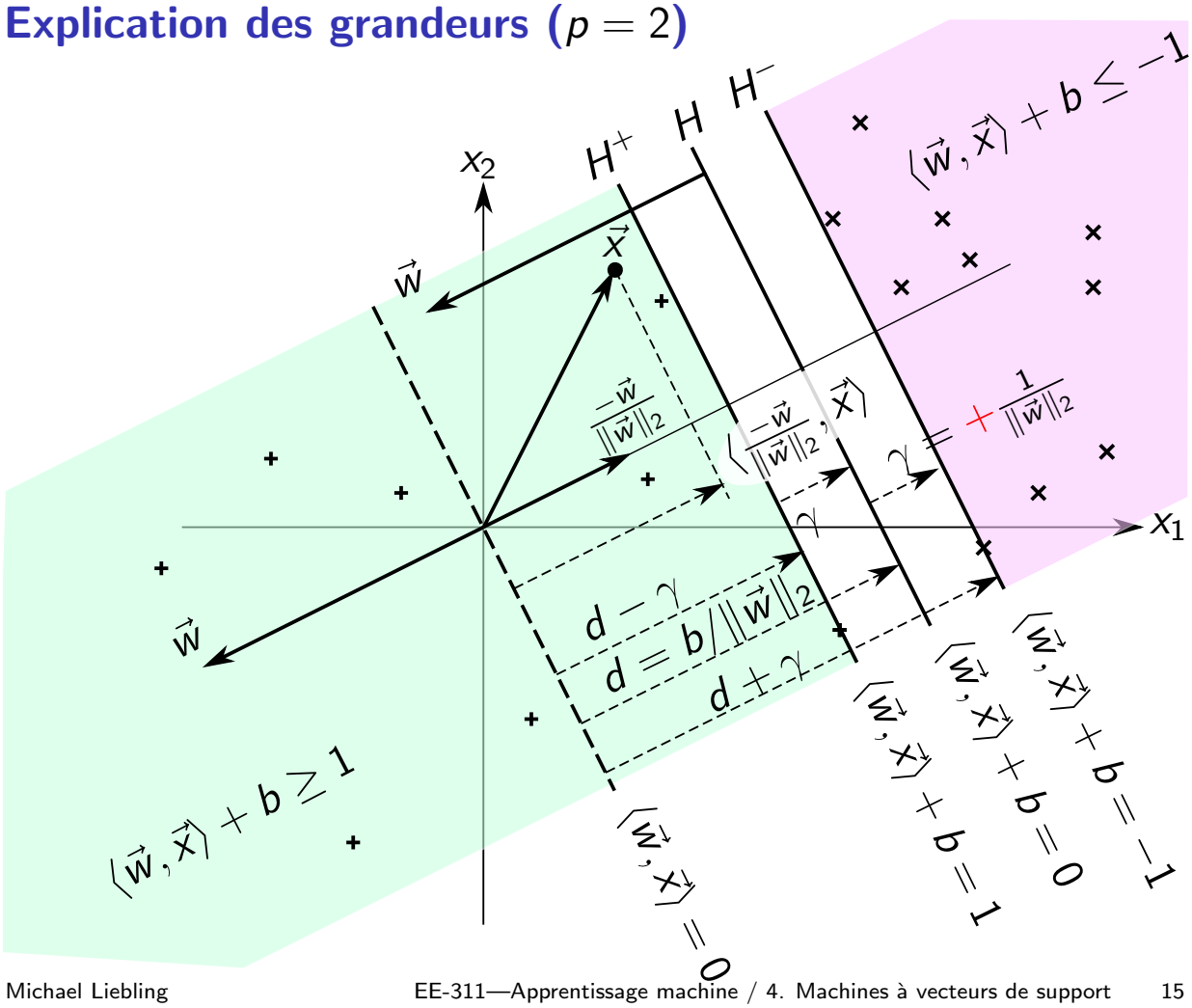
$$\langle \vec{x}, \vec{w} \rangle \geq -d\|\vec{w}\|_2 + \gamma\|\vec{w}\|_2$$

$$\langle \vec{x}, \vec{w} \rangle + \underbrace{d\|\vec{w}\|_2}_{b} \geq \underbrace{\gamma\|\vec{w}\|_2}_{1}$$

$$\langle \vec{x}, \vec{w} \rangle + b \geq 1$$

$$d = \frac{b}{\|\vec{w}\|_2} \quad \gamma = \frac{1}{\|\vec{w}\|_2}$$

Explication des grandeurs ($p = 2$)



Interprétation

Les observations positives vérifient :

$$\langle \vec{w}, \vec{x} \rangle + b \geq 1$$

Les observations négatives vérifient :

$$\langle \vec{w}, \vec{x} \rangle + b \leq -1$$

Pour le jeu d'entraînement, on a alors :

$$\left(\langle \vec{w}, \vec{x}^i \rangle + b \right) y^i \geq 1$$

On a égalité pour les vecteurs de support.
Preuve : on doit considérer les deux cas possibles
si $y^i = 1$ on a $\langle \vec{w}, \vec{x}^i \rangle + b \geq 1$ et
si $y^i = -1$ on a $\langle \vec{w}, \vec{x}^i \rangle + b \leq -1$, vérifiant la relation dans les deux cas.

Formulation primale de la SVM à marge rigide

Nous cherchons à maximiser $\frac{1}{\|\vec{w}\|_2}$ sous les n contraintes $\left(\langle \vec{w}, \vec{x}^i \rangle + b \right) y^i \geq 1$. Soit :

Définition 10.5 (Formulation primale de la SVM à marge rigide)

On appelle SVM à marge rigide le problème d'optimisation suivant :

$$\underset{\vec{w} \in \mathbb{R}^p, b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2} \|\vec{w}\|_2^2 \text{ t.q. } \left(\langle \vec{w}, \vec{x}^i \rangle + b \right) y^i \geq 1, \; i = 1, \ldots, n.$$

Supposons $\vec{w}^*$, b^* solutions du problème ci-dessus ; la fonction de décision est alors donnée par

$$f(\vec{x}) = \langle \vec{w}^*, \vec{x} \rangle + b^*.$$

Note : problème d'optimisation convexe sous n contraintes (une par point du jeu d'entraînement)

Formulation duale (de la SVM à marge rigide)

Théorème 10.1 (Formulation duale de la SVM à marge rigide) Le problème défini dans le slide précédent est équivalent à :

$$\max_{\vec{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{n} \alpha_i \alpha_\ell y^i y^\ell \langle \vec{x}^i, \vec{x}^\ell \rangle$$

$$\text{t.q.} \sum_{i=1}^{n} \alpha_i y^i = 0 \text{ et } \alpha_i \geq 0, i = 1, \ldots, n.$$

Si $\vec{\alpha}^*$ (=multiplicateurs de Lagrange) est solution du problème dual :

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i^* y^i \vec{x}^i$$

$$b^* = 1 - \min_{i:\, y^i=+1} \langle \vec{w}^*, \vec{x}^i \rangle \text{ (le plus proche de l'hyperplan } H),$$

et la fonction de décision est alors donnée par

$$f(\vec{x}) = \sum_{i=1}^{n} \alpha_i^* y^i \langle \vec{x}^i, \vec{x} \rangle + b^*.$$

Complexité algorithmique

Complexité et dimensions

- La formulation primale de la SVM est un problème d'optimisation en $p + 1$ dimensions
- La formulation duale est un problème d'optimisation en n dimensions.

Implications pratiques :

- peu de données et beaucoup de variables $\Rightarrow$ on préférera la formulation duale
- beaucoup de données peu de variables $\Rightarrow$ on préférera résoudre le problème primal.

Interprétation géométrique des α_i^* (marge rigide)

Pour caractériser la relation entre $\vec{\alpha}^*$ et $(\vec{w}^*, b^*)$, écrivons :

$$\phi(\vec{w}) = \frac{1}{2}\|\vec{w}\|_2^2 \qquad \Rightarrow \qquad \underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}}{\text{argmin}} \ \phi(\vec{w}) \qquad \leftarrow \text{primal}$$

$$g_i(\vec{w}, b) = y^i\left(\langle \vec{w}, \vec{x}^i\rangle + b\right) - 1 \qquad \text{t.q. } g_i(\vec{w}, b) \geq 0$$

Si $\vec{w}^*$, b solution du problème primal et α_i^* solution du problème dual, on a une "condition d'écart complémentaire" qui dit que :

$$\alpha_i^* g_i(\vec{w}^*, b^*) = 0 \quad \text{pour tout } 1 \leq i \leq n.$$

Deux cas sont possibles pour chacune des observations i :

1. $\alpha_i^* = 0$: le minimiseur de ϕ vérifie la contrainte et $g_i(\vec{w}^*, b^*) > 0$, i.e. $\vec{x}^i$ est à l'extérieur des hyperplans H_+ ou H_- ;
2. $\alpha_i^* > 0$: contrainte vérifiée en bordure de zone de faisabilité, i.e. quand $g_i(\vec{w}^*, b^*) = 0$ et $\vec{x}^i$ est un vecteur de support.

Ainsi : *les vecteurs de support sont les observations $\vec{x}^i$ du jeu de données correspondant aux multiplicateurs de Lagrange α_i^* non nuls.*

Le cas linéairement non séparable : SVM à marge souple

En pratique, les données ne sont généralement pas linéairement séparables !

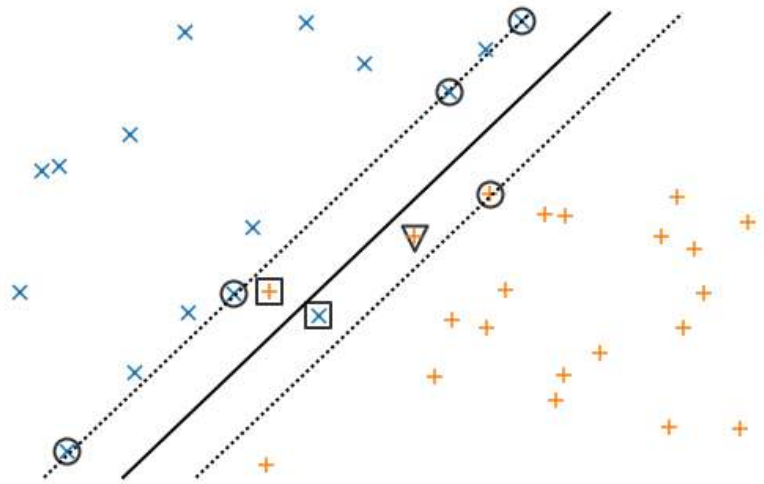


FIGURE 10.3 – Aucun classifieur linéaire ne peut séparer parfaitement ces données. Les observations marquées d'un carré sont des erreurs de classification. L'observation marquée d'un triangle est correctement classifiée mais est située à l'intérieur de la zone d'indécision. Si elle était à sa frontière, autrement dit, si elle était vecteur de support, la marge serait beaucoup plus étroite. Azencott

Formulation de la SVM à marge souple

But : trouver un compromis entre les erreurs de classification (sur le jeu d'entraînement) et la taille de la marge.
Idée : Minimiser l'inverse du carré de la marge $\|\vec{w}\|_2^2$ (comme cas rigide) et en plus un terme d'erreur pour pénaliser les instances où la classification de points du jeu d'entraînement sont erronées :

$$C \times \sum_{i=1}^{n} L\left(f(\vec{x}^i), y^i\right)$$

où $C \in \mathbb{R}^+$ est un hyperparamètre de la SVM et L la fonction de coût :

$$\underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n} L\left(\langle\vec{w}, \vec{x}^i\rangle + b, y^i\right)$$

C permet d'ajuster l'importance relative de marge et des erreurs du modèle sur le jeu d'entraînement
$\Rightarrow$ confère de la *souplesse* à la marge.

SVM à marge souple

Définition 10.6 (SVM à marge souple) On appelle SVM à marge souple la solution du problème d'optimisation suivant :

$$\underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n} \left[1 - y^i f\left(\vec{x}^i\right)\right]_+ .$$

Caractéristiques :

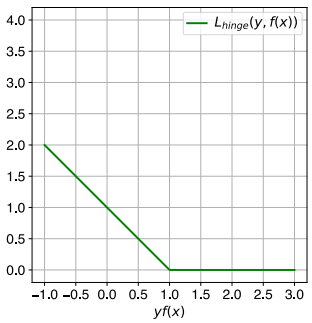
Autant que possible, on veut que toute observation $\vec{x}$ d'étiquette y soit située à l'extérieur de la zone d'indécision, i.e.

$$y^i f\left(\vec{x}^i\right) \geq 1$$

ce qui amène au choix de l'erreur hinge (voir rappel dans les slides suivants) comme fonction de coût.

(Rappel) Erreur hinge pour la classification binaire

Définition 2.12 On appelle fonction d'erreur hinge, ou hinge loss, la fonction



$$L_{\text{hinge}} : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}$$

$$y, f(\vec{x}) \mapsto \begin{cases} 0 & \text{si } yf(\vec{x}) \geq 1 \\ 1 - yf(\vec{x}) & \text{sinon} \end{cases}$$

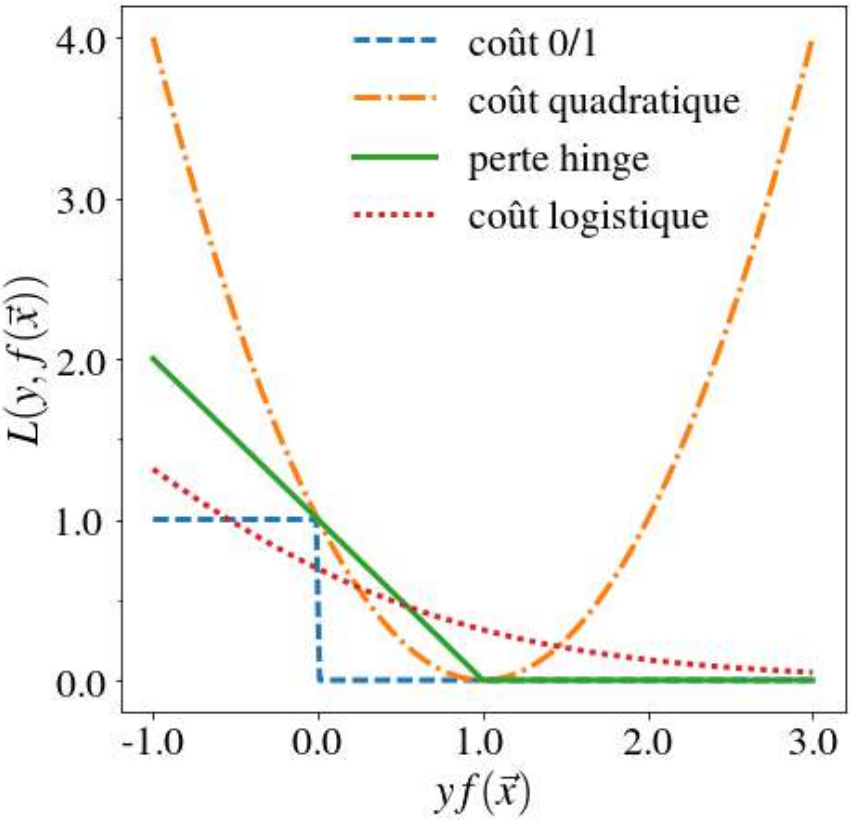
Notations équivalentes :

$$L_{\text{hinge}}(y, f(\vec{x})) = \max(0, 1 - yf(\vec{x})) = [1 - yf(\vec{x})]_+$$

Remarques

- pour une classification parfaite (quand $\mathcal{Y} = \{-1, 1\}$) on a $yf(\vec{x}) = 1$
- Fonction coût est d'autant plus grande que $yf(\vec{x})$ s'éloigne de 1 à gauche
- On considère qu'il n'y a pas d'erreur si $yf(\vec{x}) > 1$
- hinge = charnière ; aspect de coude

Rappel : fonctions de perte pour la classification binaire



Azencott

FIGURE 2.4 – Fonctions de perte pour la classification binaire.

Formulation primale de la SVM à marge souple

Définition 10.7 (Formulation primale de la SVM à marge souple) En introduisant une variable d'ajustement (ou variable d'écart—slack variable) $\xi_i = \left[1 - y^i f\left(\vec{x}^i\right)\right]_+$ pour chaque observation du jeu d'entraînement, le problème d'optimisation précédent est équivalent à

$$\underset{\vec{w}\in\mathbb{R}^p, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$$

t.q.

$$y^i\left(\langle\vec{w}, \vec{x}^i\rangle + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n$$

Note : problème d'optimisation convexe sous $2n$ contraintes (toutes affines)

Formulation duale de la SVM à marge souple

Théorème 10.2 (Formulation duale de la SVM à marge souple) La formulation primale du problème de SVM à marge souple est équivalent au problème

$$\max_{\vec{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}\alpha_i\alpha_\ell y^i y^\ell \langle\vec{x}^i, \vec{x}^\ell\rangle$$

$$\text{t.q.} \sum_{i=1}^{n}\alpha_i y^i = 0 \text{ et } 0 \leq \alpha_i \underbrace{\leq C}_{\text{NEW !}}, i = 1, \ldots, n.$$

Interprétation géométrique des α_i^* (marge souple)

Caractérisation de la relation entre $\vec{\alpha}^*$ et $(\vec{w}^*, b^*)$.
Trois cas possibles pour chaque observation i :

1. $\alpha_i^* = 0$: le minimiseur de $\frac{1}{2}\|\vec{w}\|_2^2$ vérifie la contrainte et $y^i\left(\langle \vec{w}, \vec{x}^i\rangle + b\right) > 1$, i.e. $\vec{x}^i$ est à l'extérieur de la zone d'indécision ;
2. $0 < \alpha_i^* < C$: $\vec{x}^i$ est un vecteur de support situé sur la bordure de la zone d'indécision
3. $\alpha_i^* = C$: on a $\left[1 - y^i\left(\langle \vec{w}, \vec{x}^i\rangle + b\right)\right]_+ > 0$ et $\vec{x}^i$ est du mauvais côté de la frontière d'indécision.

Relation entre $\vec{\alpha}^*$ et $(\vec{w}^*, b^*)$

Si $\vec{\alpha}^*$ est solution du problème dual, on a :

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i^* y^i \vec{x}^i$$

La fonction de décision est alors donnée par

$$f(\vec{x}) = \sum_{i=1}^{n} \alpha_i^* y^i \langle \vec{x}^i, \vec{x}\rangle + b^*.$$

Pour trouver b^*, on trouve une observation $\vec{x}^i$ sur la frontière (i.e. pour laquelle on a $0 < \alpha_i^* < C$) et on résout

$$y^i\left(\langle \vec{w}^*, \vec{x}^i\rangle + b^*\right) = 1$$

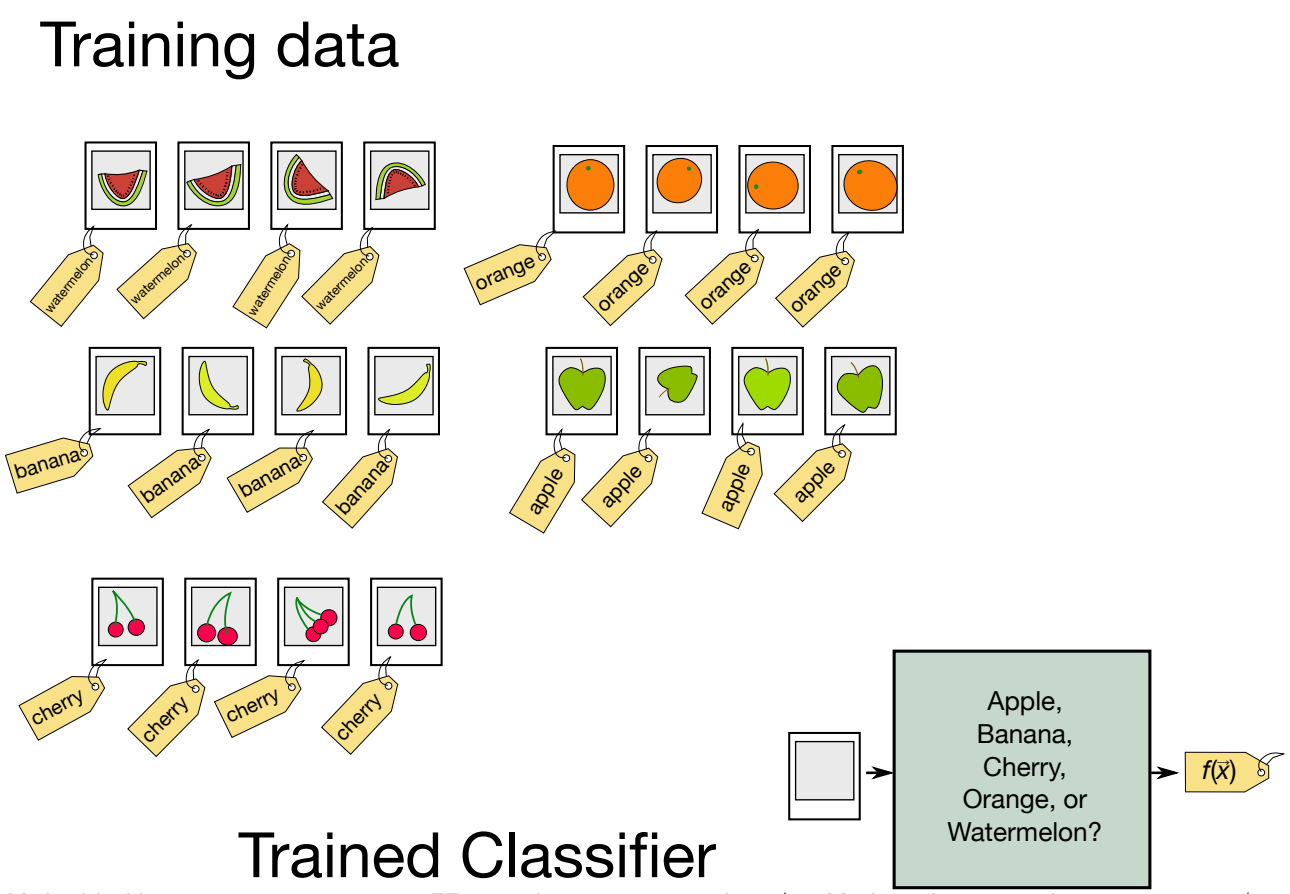
en utilisant le fait que $(y^i)^{-1} = y^i$ (comme $y = \pm 1$) :

$$b^* = y^i - \langle \vec{w}^*, \vec{x}^i\rangle$$

SVM pour classification multi-classe ?

Il est possible d'utiliser les SVMs pour construire un classificateur multi-classe, grâce à une approche une-contre-toutes ou une-contre-une.

Rappel : Classification multi-classe : comment s'y prendre ?



Rappel : Classification multi-classe avec classifieurs binaires

On peut utiliser tout algorithme de classification binaire pour résoudre un problème de classification à C classes.
Deux possibilités :

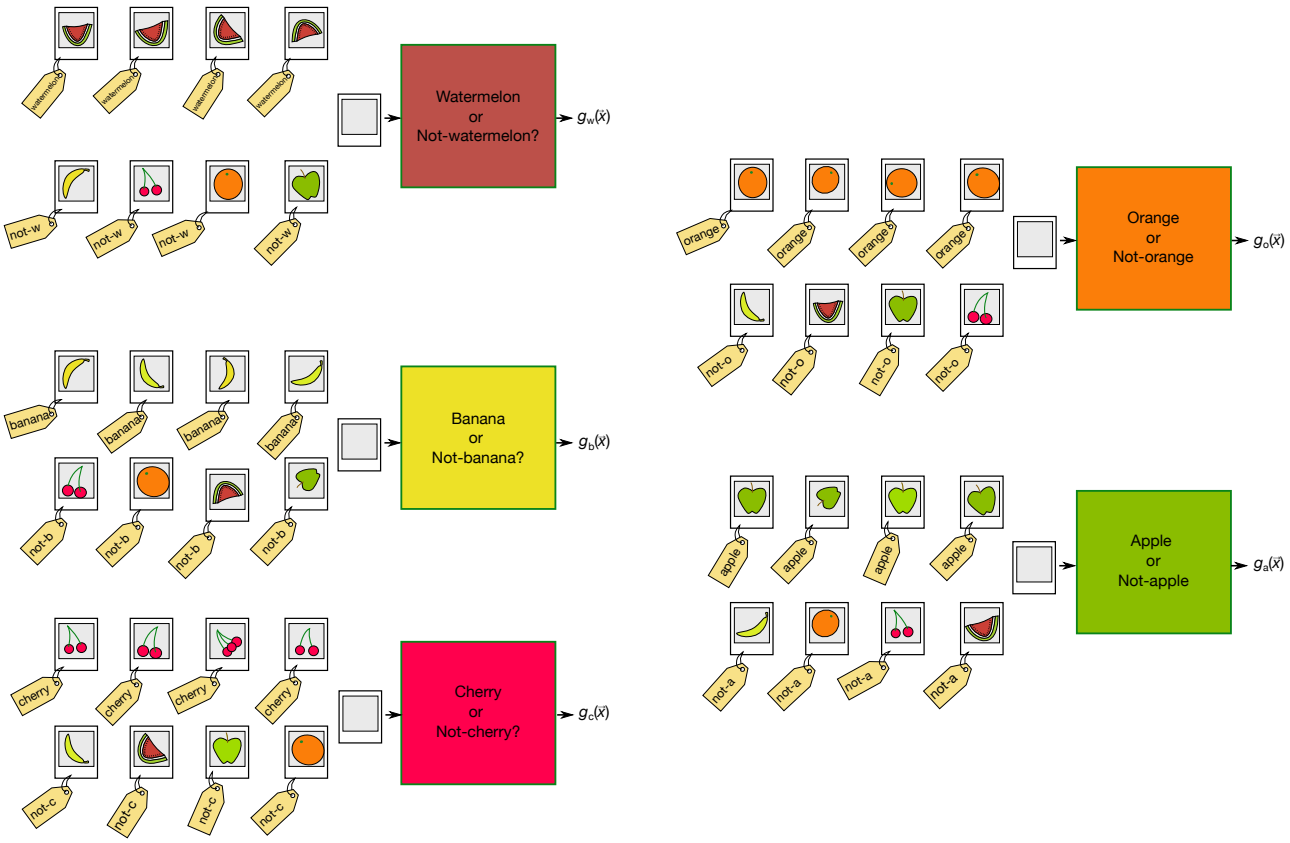
Définition 2.4 : Approche une-contre-toutes (one-versus-all)
1. Entraîner C classifieurs binaires "classe c : oui/non ?" sur l'ensemble des données d'entraînement (les exemples de la classe c sont positifs, tous les autres exemples sont négatifs)
2. Classifieur multi-classe obtenu via :

$$f(\vec{x}) = \arg \max_{c=1,\dots,C} g_c(\vec{x})$$

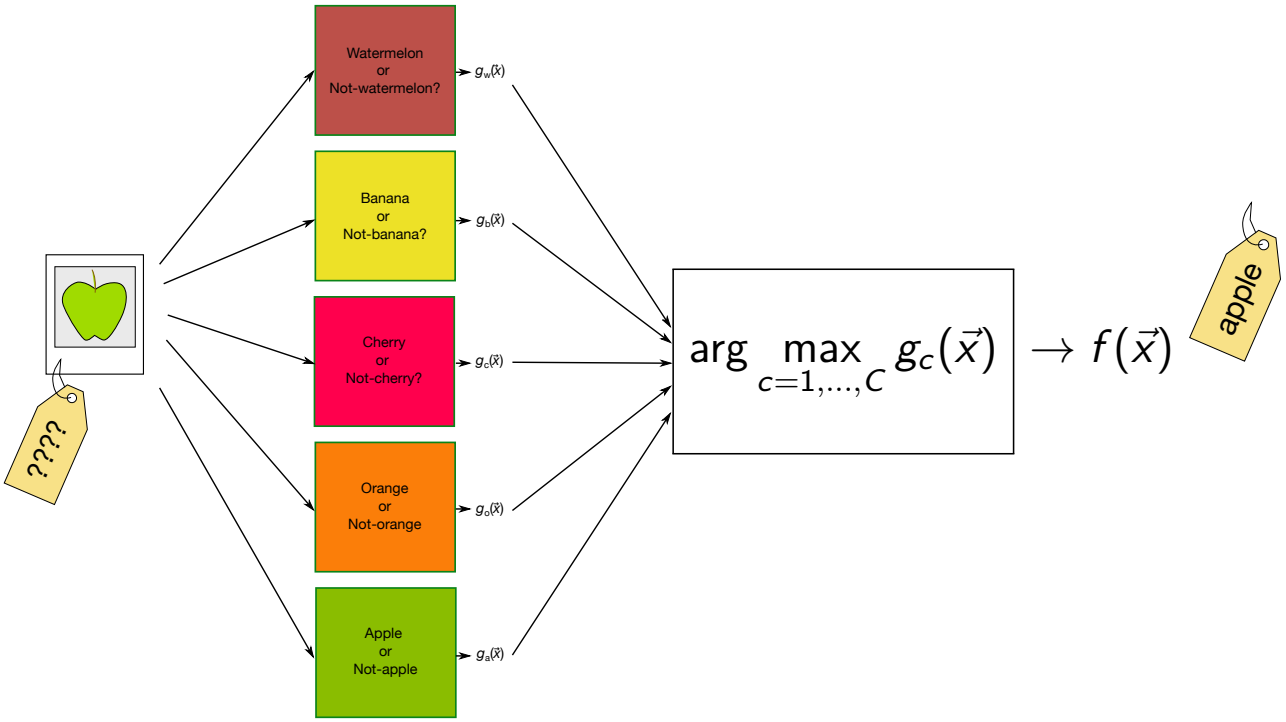
Définition 2.5 : Approche une-contre-une (one-versus-one)
1. Entraîner $C(C-1)$ classifieurs binaires "classe c : oui/non ?" sur exemples étiquetés des classes c (exemples $+$) et k (exemples -)
2. Classifieur multi-classe obtenu via :

$$f(\vec{x}) = \arg \max_{c=1,\dots,C} \left(\sum_{k \neq c} g_{ck}(\vec{x}) \right)$$

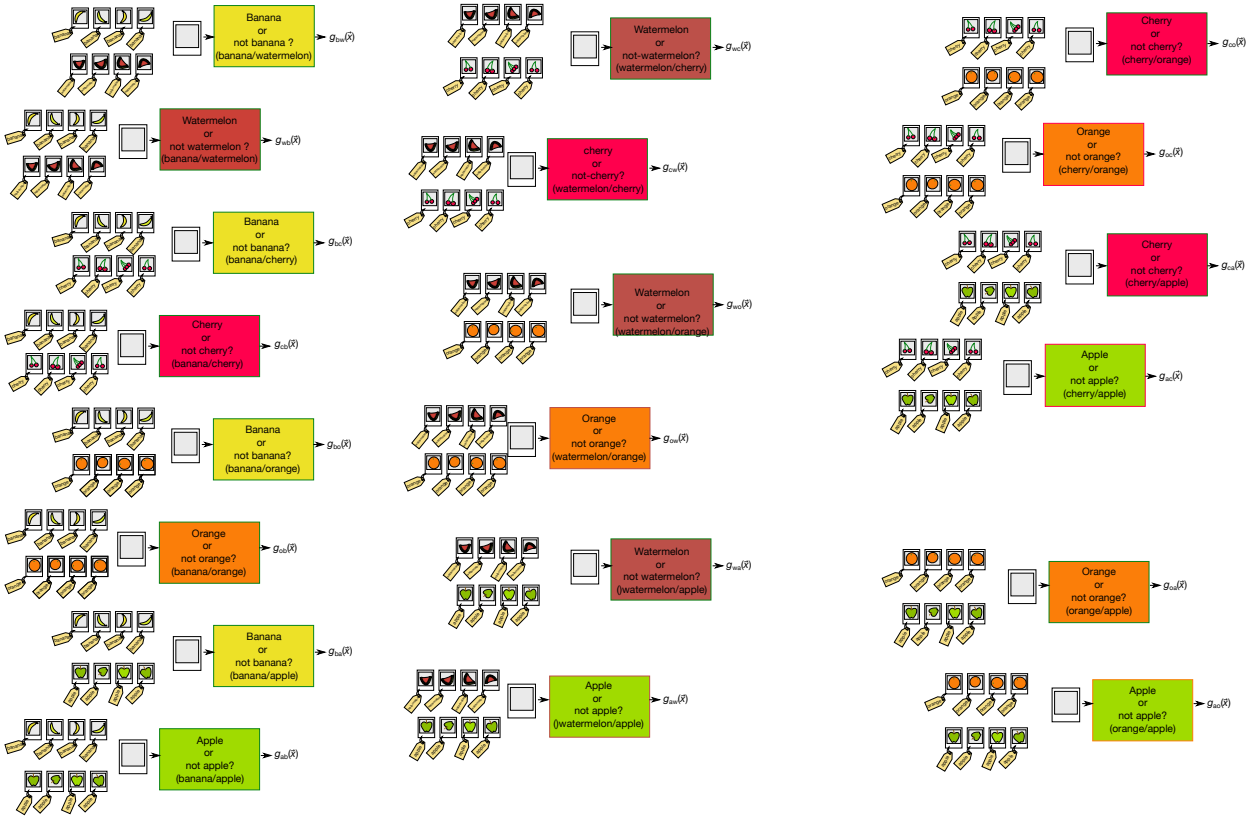
Rappel : Approche une-contre-toutes : on entraîne C classifieurs binaires (exemple : $C = 5$ classifieurs)



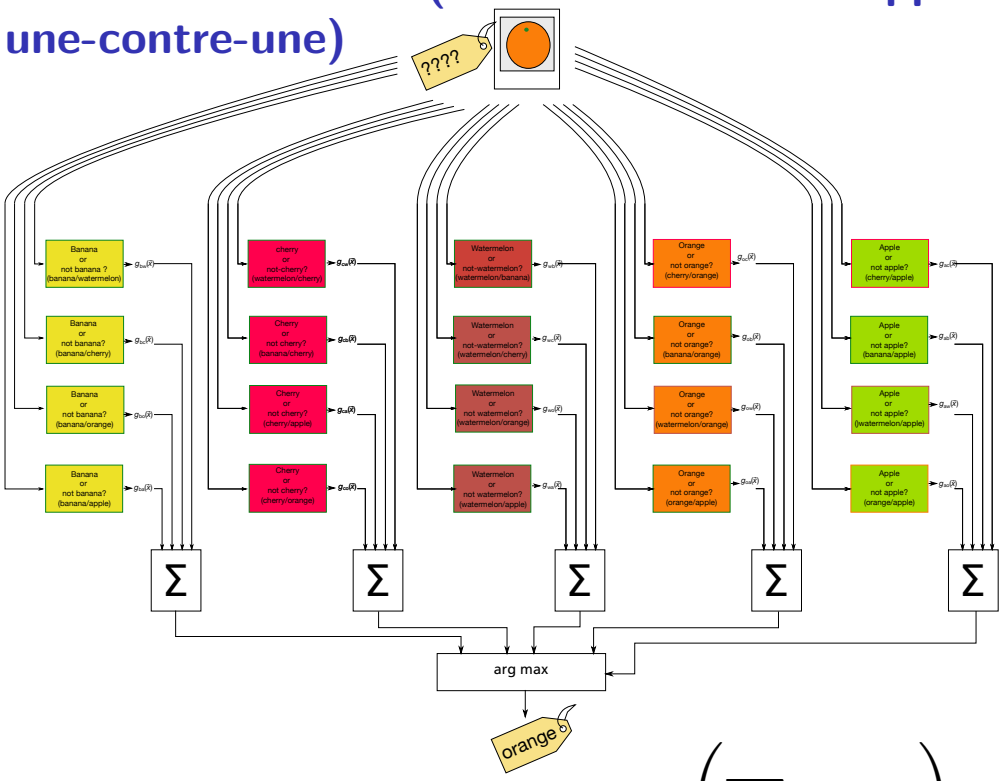
Rappel : Classifieur multi-classe construit à partir de classifieurs binaires (entraînés avec une approche une-contre-toutes)



Rappel : Approche une-contre-une : on entraîne $C(C-1)$ classifieurs binaires sur des paires de classes (exemple : $C(C-1) = 5 \times 4 = 20$ classifieurs)



Rappel : Classifieur multi-classe construit à partir de classifieurs binaires (entraînés avec une approche une-contre-une)



$$f(\vec{x}) = \arg \max_{c=1,\dots,C} \left(\sum_{k \neq c} g_{ck}(\vec{x}) \right)$$

Rappel : Une-contre-toutes ou Une-contre-une ?

Suivant la taille des données n, le nombre de classes C, le coût pour entraîner un classifieur binaire et la puissance de calcul à disposition (possibilité d'entraîner plusieurs classifieurs en parallèle), on préfèrera l'une ou l'autre approche.

Efficacité de l'entraînement (on suppose que les tailles des classes d'entraînement sont égales) :

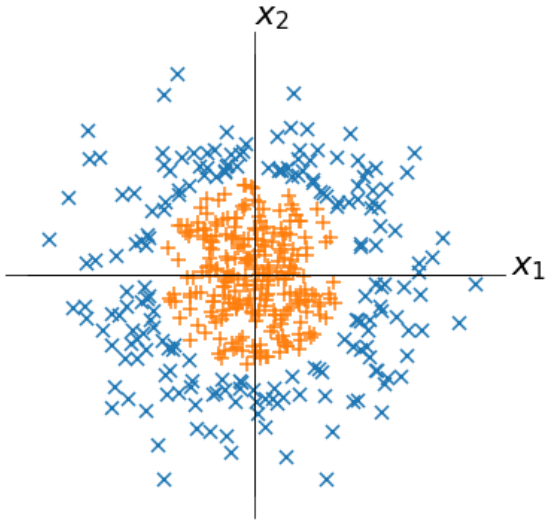
- entraîner C modèles sur n observations *ou*
- entraîner $C(C-1)$ modèles sur $2n/C$ observations ?

Qualité de l'entraînement :

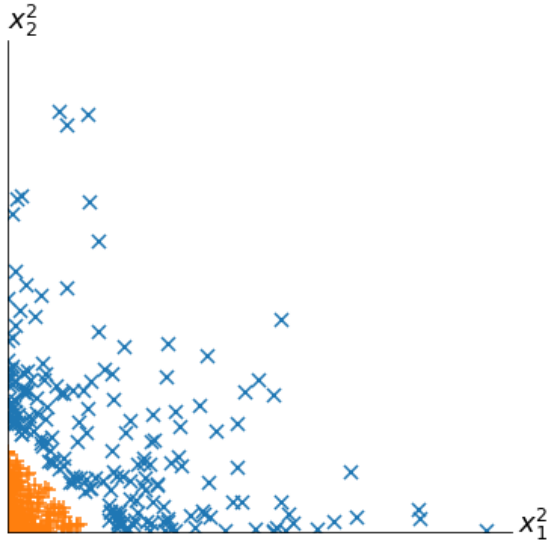
- entraîner C modèles sur n observations *ou*
- entraîner $C(C-1)$ modèles sur $2n/C$ observations ?

Cas non linéaire : SVM à noyau

Les fonctions linéaires ne sont pas toujours appropriées pour séparer les données. . .



(A) Un cercle semble bien mieux indiqué qu'une droite pour séparer ces données.



(B) Après transformation par l'application $\phi : (x_1, x_2) \mapsto (x_1^2, x_2^2)$, les données sont linéairement séparables dans l'espace de redescription.

FIGURE 10.4 – Transformer les données permet de les séparer linéairement dans un espace de redescription. Azencott

Idée : définir un espace de redescription dans lesquel la fonction de séparation est linéaire.

Espace de redescription : un exemple

Exemple : la fonction

$$\begin{aligned} f : \mathbb{R}^2 &\to \mathbb{R} \\ \vec{x} &\mapsto x_1^2 + x_2^2 - R^2 \end{aligned}$$

n'est pas linéaire en $\vec{x} = (x_1, x_2)$ mais elle est linéaire en (x_1^2, x_2^2). On peut donc définir

$$\begin{aligned} \phi : \mathbb{R}^2 &\to \mathbb{R}^2 \\ (x_1, x_2) &\mapsto \left(x_1^2, x_2^2\right) \end{aligned}$$

La fonction de décision f est linéaire en $\phi(\vec{x})$:

$$f(\vec{x}) = (\phi(\vec{x}))_1 + (\phi(\vec{x}))_2 - R^2$$

et nous pouvons l'apprendre en utilisant une SVM *sur les images des données* par l'application ϕ.

Espace de redescription : cas général

Dans le cas général, les observations sont dans un espace quelconque $\mathcal{X}$:

- $\mathcal{X} = \mathbb{R}^p$
- $\mathcal{X} =$ ensemble des chaînes de caractères sur un alphabet donné
- $\mathcal{X} =$ espace de tous les graphes
- $\mathcal{X} =$ espace de fonctions

Définition 10.8 (Espace de redescription) On appelle espace de redescription l'espace de Hilbert $\mathcal{H}$ dans lequel il est souhaitable de redécrire les données, au moyen d'une application $\phi : \mathcal{X} \to \mathcal{H}$, pour y entraîner une SVM sur les images des observations du jeu d'entraînement.

La redescription des données dans un espace de Hilbert nous permet d'utiliser un algorithme linéaire, comme la SVM à marge souple, pour résoudre un problème non linéaire.

Hilbert Spaces : Inner Product

Hilbert space : "infinite dimensional vector space" with an inner product and ... (formal definition next page)
Inner product space $\mathcal{H}$ = vector space with inner product

$\mathcal{H}$-inner product : $\langle u, v\rangle \in \mathbb{R}$ **or** $\mathbb{C}$
(i) Linearity : $\langle u, \alpha v + \beta w\rangle = \alpha\langle u, v\rangle + \beta\langle u, w\rangle \ \forall \alpha, \beta \in \mathbb{C}$, $\forall u, v, w \in \mathcal{H}$.

(ii) Conjugate Symmetry : $\langle u, v\rangle^* = \langle v, u\rangle \ \forall u, v \in \mathcal{H}$.

(iii) Positive definite : $\langle u, u\rangle > 0 \ \forall u \neq 0, u \in \mathcal{H}$.
$\langle u, u\rangle = 0 \Leftrightarrow u = 0$.
(Note : conjugate symmetry implies $\langle u, u\rangle \in \mathbb{R}$)

Induced norm

$$\|u\|_{\mathcal{H}} := \langle u, u\rangle_{\mathcal{H}}^{1/2}$$

Hilbert spaces : completeness or closedness

Completeness or closedness Every Cauchy sequence in $\mathscr{H}$ converges to a vector in $\mathscr{H}$.
Cauchy sequence $\{x_n\}$ **:** $\forall \epsilon > 0$, there exists N such that $\|x_n - x_m\| < \epsilon \ \forall n, m > N$. $\Rightarrow$ as n inreases, the points get closer and closer and converge to a limit.
Definition : a **Hilbert space** is an inner product space that is complete.
Separability : A Hilbert space is **separable** if and only if it contains a countable orthonormal basis.
Examples :
space of square-summable sequences $x \in \ell_2$.
Countable basis : $\{\delta[k-\ell]\}_{\ell \in \mathbb{Z}}$, $x[k] = \sum_{\ell \in Z} x[k]\delta[k-\ell]$
Non-countable basis :$\left\{\mathrm{e}^{\mathrm{j}\omega}\right\}_{\omega \in \mathbb{R}}$ (counter-example)

$$x[k] = \frac{1}{2\pi}\int_0^{2\pi} X\left(\mathrm{e}^{\mathrm{j}\omega}\right)\mathrm{e}^{\mathrm{j}\omega k}\,\mathrm{d}\omega$$

Separable Hilbert spaces of interest

$\mathbb{R}^N$ or $\mathbb{C}^N$: N-dimensional Euclidean space
Real or complex-valued vector $\mathbf{u} = (u_1, \ldots, u_N) = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}$
Euclidean inner product : $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^{*\top}\mathbf{v}$

$$= \sum_{i=1}^{N} u_i^* v_i$$

ℓ_2 : space of square summable sequences Real or complex-valued discrete sequences : $\{x[k]\}_{k \in \mathbb{Z}}$, ℓ_2-inner product :

$$\langle x, y \rangle = \sum_{k \in \mathbb{Z}} x^*[k]y[k]$$

L_2 : space of Lebesgue square-integrable functions
Real or complex-valued functions : $f(x), x \in \mathbb{R}$
L_2-inner product :

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f^*(x)g(x)\,\mathrm{d}x$$

SVM dans l'espace de redescription

Pour entraîner une SVM sur les images de nos observations dans l'espace de redescription $\mathscr{H}$, il nous faut donc résoudre (en utilisant la formulation duale du problème de SVM à marge souple) le problème suivant (NEW) :

$$\max_{\vec{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{\ell=1}^{n} \alpha_i \alpha_\ell y^i y^\ell \langle \phi\left(\vec{x}^i\right), \phi\left(\vec{x}^\ell\right) \rangle_{\mathscr{H}}$$

$$\text{t.q.} \sum_{i=1}^{n} \alpha_i y^i = 0 \text{ et } 0 \leq \alpha_i \leq C, i = 1, \ldots, n.$$

La fonction de décision sera ensuite donnée par :

$$f(\vec{x}) = \sum_{i=1}^{n} \alpha_i^* y^i \langle \phi\left(\vec{x}^i\right), \phi\left(\vec{x}\right) \rangle_{\mathscr{H}} + b^*.$$

SVM à noyau

Comme les images des observations obtenues par la transformation ϕ apparaissent uniquement dans des produits scalaires sur $\mathscr{H}$, nous pouvons remplacer ceux-ci avec la fonction suivante appelée **noyau** :

$$\begin{aligned} k : \mathscr{X} \times \mathscr{X} &\to \mathbb{R} \\ \vec{x}, \vec{x}' &\mapsto \langle \phi\left(\vec{x}\right), \phi\left(\vec{x}'\right) \rangle_{\mathscr{H}} \end{aligned}$$

Définition 10.9 (SVM à noyau) On appelle SVM à noyau la solution du problème d'optimisation suivant :

$$\max_{\vec{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{\ell=1}^{n} \alpha_i \alpha_\ell y^i y^\ell k\left(\vec{x}^i, \vec{x}^\ell\right)$$

$$\text{t.q.} \sum_{i=1}^{n} \alpha_i y^i = 0 \text{ et } 0 \leq \alpha_i \leq C, i = 1, \ldots, n.$$

La fonction de décision sera ensuite donnée par :

$$f(\vec{x}) = \sum_{i=1}^{n} \alpha_i^* y^i k\left(\vec{x}^i, \vec{x}\right) + b^*.$$

Astuce du noyau English : kernel trick

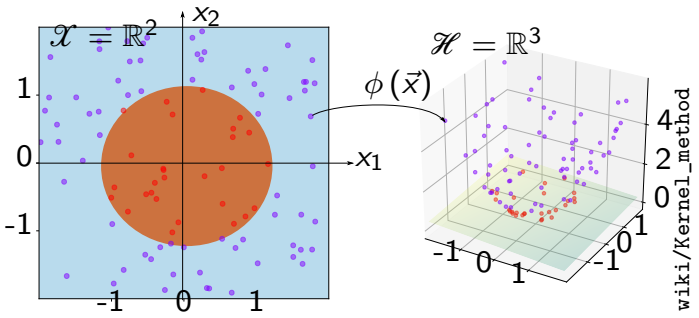
Que ce soit pour entraîner la SVM ou pour l'appliquer, nous n'avons pas besoin de connaître ϕ explicitement, mais il nous suffit de connaître le noyau k.

Cela signifie que nous n'avons pas besoin de faire de calcul dans $\mathscr{H}$, qui est généralement de très grande dimension : c'est ce que l'on appelle l'astuce du noyau.

L'astuce du noyau s'applique de manière générale à d'autres algorithmes d'apprentissage linéaires, comme la régression ridge, l'ACP ou encore la méthode des K-moyennes.

Illustration : effectuer une SVM dans un espace où c'est possible (car les données transformées y sont séparables)

Pas de séparabilité en $\mathscr{X} = \mathbb{R}^2$, mais hyperplan existe si transformation dans un espace $\mathscr{H} = \mathbb{R}^3$



Espace de redescription :

$$\phi : \mathscr{X} = \mathbb{R}^2 \to \mathscr{H} = \mathbb{R}^3$$
$$(x_1, x_2) \mapsto \left(x_1,\ x_2,\ x_1^2 + x_2^2\right)$$

Noyau (produit intérieur dans $\mathscr{H}$ *sans* transformation explicite) :

$$k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$$
$$\vec{x}, \vec{x}' \mapsto \langle \phi(\vec{x}), \phi(\vec{x}') \rangle_{\mathscr{H}} = \langle \vec{x}, \vec{x}' \rangle_{\mathscr{X}} + \|\vec{x}\|_{\mathscr{X}}^2 \|\vec{x}'\|_{\mathscr{X}}^2$$

Définition du noyau (english : kernel)

Définition 10.10 (Noyau) Nous appelons noyau toute fonction k de deux variables s'écrivant sous la forme d'un produit scalaire des images dans un espace de Hilbert de ses variables. Ainsi, un noyau est une fonction continue, symétrique, et semi-définie positive :

$$\forall N \in \mathbb{N}, \forall \left(\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^N\right) \in \mathcal{X}^N \text{ et } \left(a_1, a_2, \ldots, a_N\right) \in \mathbb{R}^N,$$
$$\sum_{i=1}^{N} \sum_{\ell=1}^{N} a_i a_\ell k\left(\vec{x}^i, \vec{x}^\ell\right) \geq 0.$$

Définition 10.11 (Matrice de Gram) Étant données n observations $(\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n) \in \mathcal{X}^n$ et un noyau k sur $\mathcal{X}$, on appelle matrice de Gram de ces observations la matrice $K \in \mathbb{R}^{n \times n}$ telle que

$$K_{i\ell} = k\left(\vec{x}^i, \vec{x}^\ell\right)$$

Cette matrice est semi-définie positive.

Théorème de Moore-Aronszajn et interprétation intuitive

Pour toute fonction symétrique semi-définie positive $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, il existe un espace de Hilbert $\mathcal{F}$ et une application $\psi : \mathcal{X} \to \mathcal{F}$ telle que pour tout $\vec{x}, \vec{x}' \in \mathcal{X} \times \mathcal{X}$ on a

$$\kappa\left(\vec{x}, \vec{x}'\right) = \left\langle \psi\left(\vec{x}\right), \psi\left(\vec{x}'\right) \right\rangle_{\mathcal{F}}$$

Intuitivement, un noyau peut être interprété comme un produit scalaire sur un espace de Hilbert, autrement dit, comme une fonction qui mesure la similarité entre deux objets de $\mathcal{X}$. Ainsi, on peut définir des noyaux en construisant une similarité entre objets, puis en vérifiant qu'elle est semi-définie positive.

Noyaux pour vecteurs réels

Quand $\mathcal{X} = \mathbb{R}^p$, le théorème de Moore-Aronszajn nous permet de définir les noyaux suivants.

Définition 10.12 (Noyau quadratique) on appelle *noyau quadratique* le noyau défini par

$$\kappa(\vec{x}, \vec{x}') = \left(\langle \vec{x}, \vec{x}' \rangle + c\right)^2, \quad c \in \mathbb{R}^+$$

En comparaison, l'application ϕ correspondant à ce noyau est :

$$\phi : \vec{x} \mapsto \Big(x_1^2, \ldots, x_p^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_p, \ldots, \sqrt{2}x_{p-1}x_p, \ldots, \sqrt{2c}x_1, \ldots, \sqrt{2c}x_p, c\Big)$$

Comme ϕ a valeur dans un espace de $2p + \frac{p(p-1)}{2} + 1$ dimensions, utiliser κ et l'astuce du noyau sera plus efficace que de calculer les images des observations par ϕ avant de leur appliquer une SVM.

Autres noyaux réels

Définition 10.13 (Noyau polynomial) On appelle noyau polynomial de degré $d \in \mathbb{N}$ le noyau défini par

$$\kappa(\vec{x}, \vec{x}') = \left(\langle \vec{x}, \vec{x}' \rangle + c\right)^d, \quad c \in \mathbb{R}^+$$

Note : ce noyau correspond à un espace de redescription comptant autant de dimensions qu'il existe de monômes de p variables de degré inférieur ou égal à d, soit $\binom{p+d}{d}$.

Définition 10.14 (Noyau radial gaussien) On appelle noyau radial gaussien, ou noyau RBF (pour Radial Basis Function),de bande passante $\sigma > 0$ le noyau défini par

$$\kappa(\vec{x}, \vec{x}') = \exp\left(-\frac{\|\vec{x} - \vec{x}'\|^2}{2\sigma^2}\right).$$

Ce noyau correspond à un espace de redescription de dimension infinie (!). En effet, en utilisant le développement en série entière de la fonction exponentielle on aurait une infinité de termes.

Résumé

- SVM, classification binaire supervisée, avec fonction de décision linéaire
- Machines à vecteurs de support à marge rigide : les données sont séparables
- Machines à vecteurs de support à marge souple : les données ne sont pas séparables
- Fonctions de décision non-linéaires peuvent être considérées grâce à un espace de redescription
- SVM à noyau : l'astuce du noyau permet de réduire la complexité des calculs en ne considérant que les produits scalaires de variables redescrites (et non les images elles-mêmes, qui peuvent être de haute dimension)
- Présentation de quelques noyaux fréquemment utilisés
- Noyaux existent aussi pour des espaces non numériques (lettres, etc.)

Guide de lecture pour ce cours

Chloé-Agathe Azencott "Introduction au Machine Learning",
Dunod, 2019, ISBN 978-210-080153-4
Chapitre 10 : Machines à vecteurs de support et méthodes à noyaux