Notes sur la machine à vecteur de support

Maximilien

Juin 2020

Les notions MIASHS (dans le désordre). espace vectoriel, $\mathbb{R}^d$, espace affine ?, norme, optimisation, fonction, Lagrangien (généralisation avec contrainte d'inégalité), produit scalaire, projection orthogonale, dérivation, hyperplan, connexité ?, applications et formes linéaires, noyau.

1 Introduction

Intuitivement, on cherche une fonction $h : \mathcal{X} \mapsto \mathcal{Y}$ où $\mathcal{X} \subset \mathbb{R}^n$ représente par exemple l'ensemble des photos de chiens et de chats et $\mathcal{Y} = \{0, 1\}$ avec $0 =$ chien et $1 =$ chat. Pour cela, nous disposons d'un jeu de données $\mathcal{D} = \{(x_i, y_i)\}_{i \leq N}$ d'exemples d'apprentissages. On parle d'apprentissage car nous allons "chercher" h via un algorithme d'optimisation en nous appuyant sur ces exemples d'apprentissage. Cette stratégie s'appelle "minimisation du risque empirique" dans le sens où les performances de notre fonction h seront calculées de manière empirique sur un jeu de données $\mathcal{D}$. De plus, on souhaite trouver une "bonne" fonction dans le sens où elle ferait "peu" d'erreurs en espérance sur des données qui n'ont pas été vues lors de l'apprentissage. Ce document ne traite pas de ces notions mais de la construction du modèle dit "machine à vecteur de support".

Restreignons-nous dans un premier temps aux classifieurs linéaires. Il s'agit de fonctions qui séparent $\mathcal{X}$ en deux parties connexes sensées correspondre aux deux classe 0 et 1 de $\mathcal{Y}$ via un hyperplan.

Prenons $\boldsymbol{w} \in \mathbb{R}^n \backslash \{0\}$ – on utilise la lettre w pour *weights*, c'est-à-dire pour les poids ou paramètres de notre modèle. La fonction suivante permet ainsi de classer les points qui sont d'un côté de l'hyperplan comme appartenant à la classe 1 et de l'autre comme appartenant à la classe -1 (ou 0) :

$$\begin{aligned} h_{\boldsymbol{w}} : \mathcal{X} &\mapsto \mathcal{Y} = \{-1, +1\} \\ \boldsymbol{x} &\to \text{sign}(\langle \boldsymbol{x}, \boldsymbol{w} \rangle). \end{aligned}$$

La fonction sign prend la valeur $+1$ si le produit scalaire est positif et -1 sinon – on peut supposer l'ensemble des points appartenant à l'hyperplan (produit scalaire nul) comme étant de mesure nulle et ignorer cette situation. En pratique, nous sommes plus généralement dans le cadre d'un espace affine et nous avons un "biais" (+b). Afin de fluidifier la lecture, nous n'afficherons pas le biais $+b$ dans nos formules bien qu'il soit presque sûrement différent de 0.

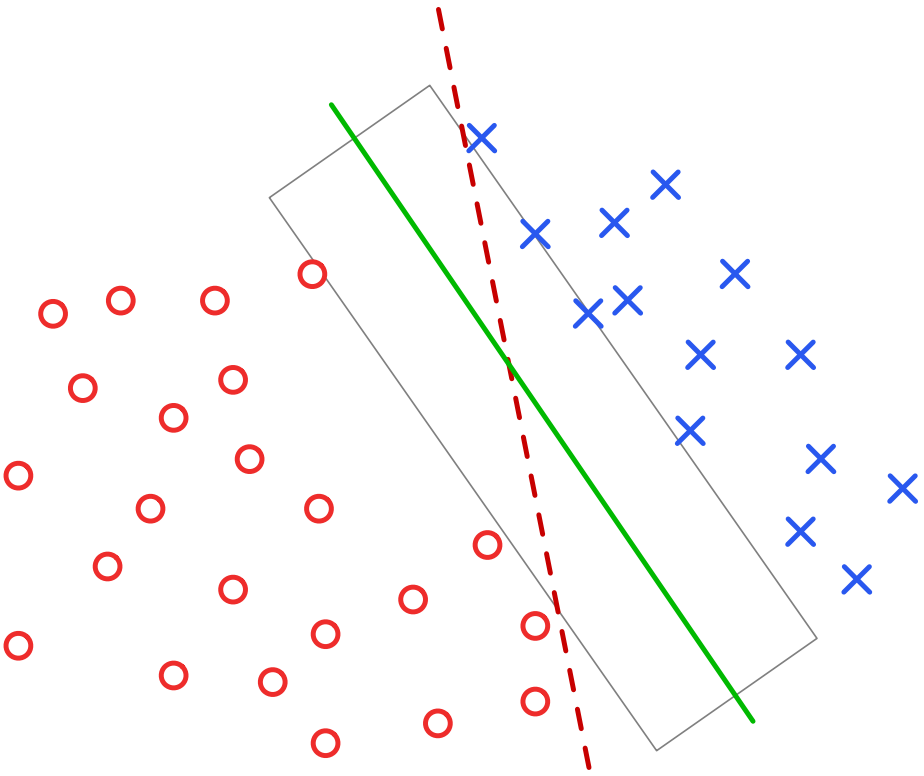

FIGURE 1 – Un problème de classification dans $\mathbb{R}^2$ avec une séparation qui maximise la marge et une séparation sans contrainte.

On s'attend intuitivement à ce que deux "points" appartenant à la même classe soient plus proches l'un de l'autre que d'un point de la classe opposée (e.g. deux photos de chiens se ressembleront peut-être plus qu'une photo de chien et une photo de chat). L'hyperplan défini par $\boldsymbol{w}$ doit être tel que de petites perturbations de nos données ne doivent pas les faire passer de l'autre côté. Cela amène à la notion de marge. Notre hyperplan doit être entouré d'une marge la plus grande possible. La figure 1 illustre cette idée. On parle de *max-margin classifier* lorsqu'on fait référence à un "classifieur"[1]. optimisé de manière à trouver $\boldsymbol{w}$ tel que la marge est maximisée.

1. Un classifieur est une fonction comme décrite précédemment dont l'espace d'arrivée est discret et non ordonné.

2 Quelques caractéristiques marrantes pour les étudiants ? ?

Soit $\boldsymbol{w} \in \mathbb{R}^n$ tel que $\|\boldsymbol{w}\| = 1$. L'ensemble des points sur l'hyperplan de vecteur normal $\boldsymbol{w}$ est donné par $\{\boldsymbol{v} : \langle \boldsymbol{v}, \boldsymbol{w} \rangle = 0\}$. La projection orthogonale de $\boldsymbol{x} \in \mathcal{X}$ sur l'hyperplan est donnée par :

$$\boldsymbol{x}_\perp = \boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{w} \rangle \boldsymbol{w}.$$

Démonstration.

$$\langle \boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{w} \rangle \boldsymbol{w}, \boldsymbol{w} \rangle = \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \langle \boldsymbol{x}, \boldsymbol{w} \rangle \langle \boldsymbol{w}, \boldsymbol{w} \rangle = 0.$$

□

La distance de $\boldsymbol{x}$ à l'hyperplan est la norme du vecteur différence entre $\boldsymbol{x}$ et son projeté orthogonal :

$$\|\boldsymbol{x}_\perp - \boldsymbol{x}\| = \|\langle \boldsymbol{x}, \boldsymbol{w} \rangle \boldsymbol{w}\| = |\langle \boldsymbol{x}, \boldsymbol{w} \rangle|.$$

On peut montrer que $\|\boldsymbol{x}_\perp - \boldsymbol{x}\|$ est bien la plus courte distance via le théorème de Pythagore (et la forme quadratique associée au produit scalaire $q(x) = \langle x, x \rangle$).

3 Le primal

Intuitivement, on cherche l'hyperplan tel que la plus petite distance des points de notre jeu de données à ce dernier soit la plus grande possible (et en gardant les points associés au même label du même côté). On dit d'ailleurs qu'un problème est séparable linéairement s'il existe un hyperplan qui sépare parfaitement notre jeu de données (i.e. tous les points d'une même classe sont du même côté de l'hyperplan et la classe opposée de l'autre). Nous nous concentrerons ici sur les cas où le problème est séparable linéairement. Évidemment le problème peut être légèrement reformulé pour "accepter" que la solution $h_{\boldsymbol{w}}$ de l'optimisation fasse des erreurs, y-compris sur notre jeu d'apprentissage. On parle de "hard-SVM" lorsqu'on peut faire cette séparation ou de "soft-SVM" si les erreurs sont tolérées. Nous nous concentrerons ainsi sur le premier cas de figure.

Le problème du SVM est réellement la recherche d'un hyperplan. En particulier, il s'agit de la solution du problème d'optimisation suivant (rappelons que nous avons volontairement omis le biais pour fluidifier la lecture) :

$$\underset{\boldsymbol{w}, \|\boldsymbol{w}\|=1}{\operatorname{argmax}} \min_{i \leq N} y_i \langle \boldsymbol{x_i}, \boldsymbol{w} \rangle \tag{1}$$

La présence du y_i force la solution à mettre les points associés à différentes classes du bon côté de la frontière (de décision). Autrement dit, on veut trouver

$\boldsymbol{w}$ de norme 1 tel que le point le plus proche de l'hyperplan défini par $\boldsymbol{w}$ soit le plus loin possible. Le problème peut se repenser sous une forme quadratique.

Soit :

$$\begin{aligned} \boldsymbol{w_0} &= \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{w}\|^2 , \ s.t. \ \forall i \leq N, y_i \langle \boldsymbol{x_i}, \boldsymbol{w} \rangle \geq 1 \\ \boldsymbol{w} &= \boldsymbol{w_0} / \|\boldsymbol{w_0}\| \end{aligned} \tag{2}$$

$\boldsymbol{w}$ est la solution du même problème d'optimisation.

On cherche ici le plus petit vecteur $\boldsymbol{w}$ tel que la "marge" reste supérieure à 1 et on le normalise. Les deux problèmes donnent la même solution.

Démonstration. Soit $\boldsymbol{w}^\star$ une solution du premier problème et soit $\gamma^\star = \min_{i \leq N} y_i \langle \boldsymbol{x_i}, \boldsymbol{w^\star} \rangle$. $\gamma^\star$ est donc la distance du point le plus proche de l'hyperplan de vecteur normal $\boldsymbol{w^\star}$ à ce dernier. Nous avons donc

$$\forall i \leq N, \ y_i \langle \boldsymbol{x_i}, \boldsymbol{w^\star} \rangle \geq \gamma^\star$$

et de manière totalement équivalente :

$$\forall i \leq N, \ y_i \langle \boldsymbol{x_i}, \boldsymbol{w^\star} / \gamma^\star \rangle \geq 1$$

Notons ici que $\boldsymbol{w^\star}/\gamma^\star$ satisfait bien les contraintes du problèmes d'optimisation quadratique (le second problème). Ainsi, il suffit de montrer qu'il n'existe pas de meilleure solution au problème quadratique, pour que nos deux problèmes soient équivalents.

Soit $\boldsymbol{w_0}$ la solution du problème quadratique avant normalisation. On a ainsi $\|\boldsymbol{w_0}\| \leq \|\boldsymbol{w^\star / \gamma^\star}\| = 1/\gamma^\star$. L'inégalité vient du fait que $\boldsymbol{w^\star / \gamma^\star}$ satisfait les contraintes mais n'est peut-être pas la meilleure solution. Soit $\hat{\boldsymbol{w}} = \boldsymbol{w_0} / \|\boldsymbol{w_0}\|$. Nous avons $\forall i \leq N$:

$$y_i \langle \boldsymbol{x_i}, \hat{\boldsymbol{w}} \rangle = \frac{1}{\|\boldsymbol{w_0}\|} y_i \langle \boldsymbol{x_i}, \boldsymbol{w_0} \rangle \geq \frac{1}{\|\boldsymbol{w_0}\|} \geq \gamma^\star.$$

La première inégalité vient de la contrainte du problème d'optimisation quadratique et la seconde que la solution du premier problème satisfait elle-même ces contraintes. Puisque $\|\hat{\boldsymbol{w}}\| = 1$ et définit une marge au moins aussi grande que $\boldsymbol{w}^\star$ qui par définition définit la plus grande marge, les solutions sont équivalentes. □

On sent bien que l'hyperplan ne dépend que des points exactement sur "les bords" de la marge mais le problème que nous optimisons ici ne le montre pour l'instant pas. En réalité, on parle de machine à vecteur de support car la solution du problème d'optimisation est nécessairement dans l'espace vectoriel engendré par les vecteurs exactement sur les bords de la marge.

4 Le dual

Nous allons ici dériver une formation duale du problème précédent. Au-delà de l'exercice, la formulation duale apporte un certain nombre d'avantages et

notamment lorsque nous considérerons autre chose que des hyperplans pour séparer nos données. Rappelons que le problème à optimiser est le suivant :

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{w}\|^2, \ s.t. \ 1 - y_i\langle \boldsymbol{x_i}, \boldsymbol{w}\rangle \leq 0, \ \forall i \leq N.$$

Nous n'avons ici pas de contrainte d'égalité mais uniquement des contraintes d'inégalités. Formulons le Lagrangien généralisé (avec inégalité) :

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}) = \|\boldsymbol{w}\|^2 + \sum_{i=1}^{N} \alpha_i(1 - y_i\langle \boldsymbol{x_i}, \boldsymbol{w}\rangle), \ \alpha_j \geq 0, \ j \leq N$$

où α_i est un multiplicateur de Lagrange. Observons le problème suivant :

$$g(\boldsymbol{w}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq \mathbf{0}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}).$$

On observe assez intuitivement que si une contrainte n'est pas satisfaite (i.e. $\exists i \leq N, 1 - y_i\langle \boldsymbol{x_i}, \boldsymbol{w}\rangle > 0$), alors la solution est ∞. De même et assez intuitivement si les contraintes sont satisfaites de manière stricte ($<$), alors $\alpha_i = 0$ et sinon ($=$), la valeur n'importe pas car le produit fera 0 dans tous les cas. On obtient ainsi :

$$g(\boldsymbol{w}) = \begin{cases} \|\boldsymbol{w}\|^2 & \text{si les contraintes sont satisfaites,} \\ \infty & \text{sinon.} \end{cases}$$

Autrement dit, minimiser $g(\boldsymbol{w})$ revient à minimiser le problème quadratique définit par l'équation 2. Considérons donc le problème :

$$p^\star = \min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq \mathbf{0}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}),$$

où $p^\star$ est la solution du problème dit primal.

Le problème dual associé consiste à inverser l'ordre de la minimisation et de la maximisation :

$$d^\star = \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq \mathbf{0}} \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}),$$

où on peut vérifier qu'on a nécessairement $d^\star \leq p^\star$. Ainsi, dans le cadre du primal, le problème est une minimisation, mais devient une maximisation dans le dual. Le gap de dualité donne l'écart entre le primal et le dual. On parle de dualité faible si le gap est positif et de dualité forte s'il est nul. Il se trouve que si $\|\boldsymbol{w}\|^2$ et $\alpha_i(1 - y_i\langle \boldsymbol{x_i}, \boldsymbol{w}\rangle)$ sont convexes en $\boldsymbol{w}$ et α_i respectivement, et qu'il existe $\boldsymbol{w}$ tel que $1 - y_i\langle \boldsymbol{x_i}, \boldsymbol{w}\rangle < 0$ (inégalité stricte impliquant que $\boldsymbol{w}$ est dans l'enveloppe affine de l'ensemble de faisabilité), alors nous avons une dualité stricte (conditions de Slater) [2]. L'hypothèse de séparabilité linéaire implique que ces conditions sont satisfaites. On obtient donc :

2. Notons que les conditions de Karush-Kuhn-Tucker (KKT) nous permettent également d'obtenir le résultat souhaité. Les conditions KKT permettent de généraliser l'idée des multiplicateurs de Lagrange dans le cas où des contraintes d'inégalité existent.

$$p^\star = \mathcal{L}(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star) = d^\star.$$

Concentrons-nous sur l'expression duale. Pour un $\boldsymbol{\alpha}$ donné, $\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha})$ minimise $\mathcal{L}$ relativement à $\boldsymbol{w}$. Ainsi, nous avons :

$$0 = \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 2\boldsymbol{w} - \sum_i \alpha_i y_i \boldsymbol{x_i} \Leftrightarrow \boldsymbol{w} = \frac{1}{2} \sum_i \alpha_i y_i \boldsymbol{x_i}.$$

Autrement dit, nous pouvons reformuler le problème dual de la manière suivante :

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq \mathbf{0}} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle.$$

Ce problème d'optimisation est équivalent au primal et consiste à maximiser une quantité qui ne dépend plus que du produit scalaire entre les points de notre jeu de données (à un signe près s'ils appartiennent à des classes opposées).

Les coefficients $\boldsymbol{\alpha}$ calculés, nous souhaitons revenir à notre règle de décision. Pour cela il suffit de remplacer $\boldsymbol{w}$ dans la construction initiale :

$$\begin{aligned} h : \mathcal{X} &\mapsto \mathcal{Y} \\ \boldsymbol{x} &\to \text{sign}(\sum_i \alpha_i y_i \langle \boldsymbol{x_i}, \boldsymbol{x} \rangle) \end{aligned}$$

Intuitivement, le produit scalaire $\langle \cdot, \cdot \rangle$ peut être vu comme une mesure de similarité et un point sera "prédit" comme appartenant à la classe des points dont il est le plus similaire. De plus, notons que $\alpha_i = 0$ si le point n'est pas exactement sur le bord de la marge. Ainsi, les points considérés sont ceux situés exactement sur le bord de la marge.

5 Espaces de Hilbert à Noyaux Reproduisant ou l'astuce du noyau

En pratique, le problème n'est pas toujours séparable linéairement et on veut pouvoir trouver une séparation non linéaire (i.e. autre chose qu'un hyperplan). Une manière standard est de chercher des applications $\phi : \mathcal{X} \mapsto F$ et d'optimiser notre SVM dans F. En choisissant bien ϕ, on peut espérer qu'une séparation linéaire dans F correspondra bien à la frontière non-linéaire recherchée dans $\mathcal{X}$. Le problème à optimiser devient ainsi :

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq \mathbf{0}} \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{\phi(x_i)}, \boldsymbol{\phi(x_j)} \rangle.$$

Un exemple classique est la "transformation polynomiale". Dans le plan avec un polynôme de degré 2, nous avons :

$$\begin{aligned} \phi : \mathcal{X} &\mapsto F \subset \mathbb{R}^3 \\ x_1, x_2 &\to (x_1^2, \sqrt{2} x_1 x_2, x_2^2) \end{aligned}$$

Nous verrons plus bas que ce choix et notamment les coefficients (inutiles à ce stade) n'est en réalité pas anodin. Continuons d'omettre le biais de nos formules.

Quelques mots sur les RKHS. Considérons un espace de Hilbert de fonctionnelles sur X (arbitraire) qu'on notera H. On appelle fonctionnelle d'évaluation L_x sur H une forme linéaire sur H telle que $\forall f \in H$, L_x associe f à son évaluation en $x \in X$, $f(x)$. De plus, $\|L_x\|_H$ (où $\|\cdot\|_H$ est une norme d'opérateur sur H) est fini si et seulement si L_x est continue $\forall f \in H$. Dans ce cas de figure, H est appelé "espace de Hilbert à noyau reproduisant". Autrement dit, si $\exists M > 0$ tel que $|f(x)| \leq M \|f\|_H, \forall f \in H$, alors H est un RKHS et inversement. L'espace H étant de Hilbert, le théorème de réprésentation de Riesz nous dit $\exists! g \in H$ tel que $\forall f \in H$, on a $L_x(f) = \langle f, g \rangle$ (le produit scalaire est pris dans H). La fonction g est elle-même un élément de H et peut être évaluée en un point $y \in X$. Soit h telle que $L_y(x) = \langle h, x \rangle$. On a donc :

$$L_y(g) = \langle h, g \rangle = L_x(h)$$

où h et g sont des fonctionnelles d'évaluation. Notons $K_x \in H$ la représentation dans H de L_x. Nous pouvons ainsi définir une application sur $X \times X$ dans $\mathbb{R}$ appelée noyau :

$$K(x, y) = \langle K_x, K_y \rangle,$$

où le produit scalaire est pris dans H. Le noyau K est symétrique défini positif.

Transformation non linéaire et RKHS. Comme indiqué plus tôt, la frontière recherchée n'est pas nécessairement linéaire et il convient de construire des transformations non linéaires de nos données qui "linéariseraient" le problème. Nous avons appelé ces fonctions $\phi : \mathcal{X} \mapsto F$ où F est un espace de Hilbert. On appelle souvent ces espaces de "représentations" *features space*. De plus le terme représentation doit être compris dans le sens où tout $x \in \mathcal{X}$ est représenté par un élément de F. Le problème dual du SVM nous a montré qu'on pouvait résumer la formulation à des produits scalaires entre nos données. Nous pouvons alors définir le noyau suivant :

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_F.$$

Le produit scalaire $\langle \cdot, \cdot \rangle_F$ est par définition symétrique défini positif, et pour un choix adéquat de ϕ, $K(\cdot, \cdot)$ est également symétrique défini positif. De plus, notons :

$$K_x : \boldsymbol{y} \mapsto K(\boldsymbol{x}, \boldsymbol{y}) \quad \text{et} \quad K_y : \boldsymbol{x} \mapsto K(\boldsymbol{x}, \boldsymbol{y}).$$

L'ordre n'importe pas du fait de la symétrie. Ainsi, il est possible de construire un espace de Hilbert ayant la propriété reproduisante de la manière suivante. Soit H cet espace. Tout $x \in \mathcal{X}$ permet de construire une fonction $K_x \in H$. En outre, H doit également contenir toutes les combinaisons linéaires d'éléments de la forme K_x. De plus, il est possible de construire le produit scalaire $\langle \cdot, \cdot \rangle_H$

à partir de $\langle \cdot, \cdot \rangle_F$. Il restera ici à montrer/constater la complétude de H. On notera que la fonction calculée par notre SVM appartient à cet espace :

$$\sum_i \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) \in H.$$

De manière illustrée et simple, si ϕ est un endomorphisme sur $\mathbb{R}^d$, le produit scalaire sur $\mathbb{R}^d$ est un produit scalaire sur l'espace dual qui lui-même est un espace de Hilbert à noyau reproduisant.

Notons de plus que les combinaisons linéaires et produits de noyaux définissent de nouveaux noyaux.

Un exemple. Reprenons l'application suivante :

$$\begin{aligned} \phi : \mathcal{X} &\mapsto F \subset \mathbb{R}^3 \\ x_1, x_2 &\to (x_1^2, \sqrt{2} x_1 x_2, x_2^2) \end{aligned}$$

Dans le cadre du SVM, ce qui nous intéresse n'est pas la transformation en elle-même mais la valeur du produit scalaire $\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_F$. Définissons le noyau sur $\mathbb{R}^2$ suivant :

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle)^2 = x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2.$$

On retrouve le produit scalaire qu'on aurait obtenu avec ϕ. En réalité, le noyau :

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c)^n$$

est celui qui donne le produit scalaire entre deux points de notre jeu de données après une transformation polynomiale de degré n. L'astuce réside dans le fait que seul le produit scalaire dans $\mathcal{X}$ a besoin d'être calculé.

Cette astuce peut se généraliser avec des noyaux $\phi : \mathcal{X} \mapsto F$ où F est de dimension infinie (e.g. le noyau rbf).