Reconnaissance de formes par méthodes à noyaux

Paul HONEINE

Université de technologie de Troyes

— *draft* 2013 —

Table des matières

Chapitre 1

Apprentissage et méthodes à noyaux

Ce chapitre a pour objectif de présenter un cadre théorique au problème d'apprentissage statistique, qui consiste à déterminer une fonction à partir de données regroupées dans un ensemble d'apprentissage. Ce problème est mal-posé puisqu'il existe une infinité de fonctions continues qui vérifient les conditions discrètes induites par les données d'apprentissage. Ceci est par exemple le cas des problèmes de régression, où l'on cherche une fonction passant en certains points tandis qu'il en existe une infinité.

La théorie de la régularisation introduite par Tikhonov et Arsenin dans [TA77] propose une solution élégante à ce problème. Une régularisation de type Tikhonov permet de restreindre la recherche à un espace de fonctions régulières. Un type d'espace fonctionnel particulier est l'espace de Hilbert à noyau reproduisant, un concept introduit par Aronszajn dans [Aro50]. Ses propriétés sont exploitées par le Théorème de Représentation, initialement proposé pour les problèmes de régression par Kimeldorf et Wahba dans [KW71, Wah90], et récemment généralisé à d'autres problèmes d'apprentissage par Schölkopf *et coll.* dans [SHW00]. La simplicité des méthodes dites à noyaux est principalement due au coup du noyau, plus communément désigné par *kernel trick* en anglais, qui permet de transformer des algorithmes linéaires en des méthodes non-linéaires sans surcoût calculatoire considérable, sous réserve que ceux-ci puissent s'exprimer uniquement par des produits scalaires entre les données. Cette notion de non-linéarité par usage de noyau a été proposée par Aizerman *et coll.* dans [ABR64] dans le cadre d'un problème de classification, et renforcé par Vapnik dans [Vap95] avec le théorie de l'apprentissage statistique dans un contexte plus général

de classification et régression.

Dans ce chapitre, on traite d'une manière concise ces différents concepts, que l'on illustre avec certains exemples de méthodes à noyau que l'on détaillera au cours des chapitres suivants. Dans la Section 1.1, on introduit les méthodes d'apprentissage et la régularisation selon Tikhonov. Après avoir présenté succinctement les concepts de noyau reproduisant et d'espace de Hilbert associé dans la Section 1.2, on introduit les deux clés de voûte des méthodes à noyaux dans la Section 1.3 : le coup du noyau et le Théorème de Représentation. On conclut le chapitre par des exemples de méthodes à noyaux dans des problèmes d'analyse non-supervisée et supervisée.

1.1 Théorie de l'apprentissage statistique

En théorie de l'apprentissage statistique [Vap95], on cherche à déterminer un modèle qui traduit le mieux possible une relation entre les observations successives recueillies sur un système, ou encore entre ses entrées et sorties, à partir d'un ensemble d'apprentissage. On souhaite que le modèle ainsi élaboré soit généralisable à de nouvelles observations, ce qui nécessite une certaine connaissance *a priori* du comportement du système. Une telle information peut être incorporée grâce à un choix approprié d'espace d'hypothèses, ce dernier étant l'espace fonctionnel dans lequel la solution est recherchée.

1.1.1 Apprentissage statistique

Les méthodes d'apprentissage statistique peuvent être regroupées en deux classes principales : les méthodes supervisées et les méthodes non-supervisées.

Dans le cadre de l'apprentissage supervisé, on cherche la relation entre un compact $\mathcal{X}$ de $\mathbb{C}^l$, dit espace de données ou d'entrée, et un compact $\mathcal{Y}$ de $\mathbb{C}$, dit espace des réponses ou de sortie. Cette relation est décrite par la distribution de probabilité $P(x, y)$ définie pour tout couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Ainsi recherche-t-on la fonction ψ^* de $\mathcal{X}$ dans $\mathcal{Y}$ telle que $\psi^*(x)$ soit une estimation appropriée de la sortie y correspondant à la donnée x. L'optimalité de la fonction ψ^* sur toutes les fonctions ψ de $\mathcal{X}$ dans $\mathbb{C}$ est donnée par la minimisation d'une fonctionnelle de risque réelle de la forme

$$\int_{\mathcal{X}\times\mathcal{Y}} V(\psi(x), y)\, \mathrm{d}P(x, y),$$

où $\mathrm{d}P(x,y) = P(x,y)\,\mathrm{d}x\,\mathrm{d}y$, et V une fonction coût qui mesure l'erreur commise entre la sortie désirée y et la sortie estimée $\psi(x)$ pour tout couple $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Dans le cas particulier de la fonction de coût quadratique, $V(\psi(x),y) = (\psi(x)-y)^2$, la fonction minimisant cette expression est donnée par $\psi^*(x) = \int_{\mathcal{Y}} y\,\mathrm{d}P(y|x)$. Puisque $\mathcal{Y}$ est compact, on montre qu'un tel optimum existe. Toutefois, la distribution de probabilité P étant inconnue, l'optimum ne peut pas être obtenu directement. Celle-ci n'est en effet connue qu'à partir d'un ensemble fini de réalisations, appelé ensemble d'apprentissage, que l'on note $\mathcal{A}_n = \{(x_1,y_1),\ldots,(x_n,y_n)\}$ avec $(x_k,y_k) \in \mathcal{X} \times \mathcal{Y}$. En posant $P_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \delta(x-x_i)\,\delta(y-y_i)$, le problème d'optimisation se traduit par la minimisation du risque d'apprentissage, appelé aussi risque empirique, selon

$$\psi^* = \arg\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i), y_i). \tag{1.1}$$

Pour un apprentissage non-supervisé, on se contente d'un compact $\mathcal{X}$ de $\mathbb{C}^l$, dit espace de données ou des observations. On cherche alors la relation entre les éléments de cet espace, que décrit la distribution de probabilité $P(x)$ pour tout $x \in \mathcal{X}$. La fonction recherchée ψ^* est alors obtenue en résolvant un problème d'optimisation de la forme

$$\psi^* = \arg\min_{\psi} \int_{\mathcal{X}} V(\psi(x))\,\mathrm{d}P(x),$$

portant sur toutes les fonctions ψ de $\mathcal{X}$ dans $\mathbb{C}$, avec V une fonction coût donnée. N'ayant à disposition qu'un ensemble d'apprentissage fini $\mathcal{A}_n = \{x_1,\ldots,x_n\}$ de réalisations échantillonnées selon P, cette dernière est estimée par $P_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x-x_i)$. On recherche une fonction optimale en minimisant le risque empirique défini par

$$\psi^* = \arg\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i)). \tag{1.2}$$

Comme cette expression est un cas particulier de (1.1) pour des étiquettes y_i supposées constantes, on considère dans la suite le cas plus général de l'apprentissage supervisé.

Dans les deux cas, il existe une infinité de fonctions ψ^* minimisant le risque empirique, donc vérifiant soit (1.1) pour l'apprentissage supervisé, soit (1.2) pour l'apprentissage non-supervisé. Le problème est dit alors mal-posé, dans le sens où l'ensemble d'apprentissage ne permet pas une reconstruction

unique de la fonction recherchée. Pour autant, toutes les fonctions candidates ψ^* n'admettent pas les mêmes capacités en généralisation étant donné de nouvelles observations ne figurant pas dans l'ensemble d'apprentissage. On a alors recours à l'introduction d'hypothèses vis-à-vis de la fonction ψ^* recherchée afin de s'affranchir du caractère mal-posé du problème initial. Une contrainte faible et naturelle, au sens des phénomènes physiques par exemple, consiste à supposer que cette fonction est suffisamment régulière pour que de faibles variations des données produisent de légères fluctuations sur les sorties. Cette contrainte de régularité sur ψ^* permet alors d'interpréter le problème d'apprentissage comme un exercice d'approximation à partir de données bruités. D'autres contraintes, plus fortes, peuvent aussi être considérées préalablement à l'apprentissage, par exemple que le modèle recherché est linéaire ou quadratique.

1.1.2 Méthodes régularisées

Depuis les années 1960, plusieurs techniques de régularisation ont été proposées pour rendre un problème d'optimisation bien-posé, dont les régularisations d'Ivanov [Iva76], de Phillips [Phi62] et de Tikhonov [Tik63] [1]. Ces techniques offrent un cadre mathématique général pour résoudre les problèmes d'optimisation (1.1) et (1.2) en restreignant l'espace de recherche des fonctions candidates aux fonctions à faibles oscillations. On considère un espace de Hilbert $\mathcal{H}$ de fonctions de $\mathcal{X}$ dans $\mathbb{C}$, auquel appartient la fonction recherchée ψ^*. Soit $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ le produit scalaire associé à cet espace, et $\| \cdot \|_{\mathcal{H}}$ sa norme. La pénalisation d'Ivanov consiste à minimiser le risque empirique, avec une contrainte sur la complexité définie par $\|\psi\|_{\mathcal{H}} \leq \tau$ qui vise à pénaliser les solutions oscillantes. Pour mieux comprendre cette pénalisation restreignant l'espace $\mathcal{H}$ aux fonctions à normes réduites, il suffit de considérer par exemple l'espace de Banach $\mathcal{L}_1[a,b]$ des fonctions intégrables en valeur absolue sur l'intervalle $[a,b]$. La norme est alors définie par $\|\psi\|_1 = \int_a^b |\psi(x)| \, \mathrm{d}x$. Un autre type de régularisation concerne l'espace $\mathcal{L}_2[a,b]$ des fonctions d'énergie finie sur $[a,b]$, avec la norme quadratique $\|\psi\|_2^2 = \int_a^b |\psi(x)|^2 \, \mathrm{d}x$. La régularisation de Phillips [Phi62] consiste à minimiser la norme, sous contrainte que le risque empirique reste inférieur à un seuil donné. Ces modes de régularisation d'Ivanov et de Phillips sont équivalents à la régularisation de Tikhonov [Muk04].

Par l'usage de la technique des multiplicateurs de Lagrange, Tikhonov propose dans [TA77] de transformer le problème d'optimisation avec

1. Cette pénalisation est souvent connue sous le nom de Tikhonov–Phillips [MR97].

contrainte défini ci-dessus en un problème d'optimisation sans contrainte. La fonctionnelle de risque étant pénalisée, la solution est alors obtenue selon

$$\psi^* = \arg\min_{\psi\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta \,\|\psi\|^2_{\mathcal{H}},$$

où η contrôle le compromis entre les deux termes. Le premier terme représente le risque empirique qui mesure l'adéquation entre les sorties estimées et les sorties désirées. Le second terme, de pénalisation, permet d'obtenir des solutions plus régulières. Sans celui-ci, le problème serait mal-posé puisqu'il existerait alors une infinité de fonctions qui minimisent le premier terme. La théorie de la régularisation *à la Tikhonov* des méthodes d'apprentissage statistiques a connu diverses avancées récentes depuis qu'elle a été introduite dans ce contexte par Poggio et Girosi [PG90], et plus récemment avec les techniques de SVM proposées par Vapnik [Vap95]. Une généralisation de cette pénalisation a également été considérée en remplaçant le terme $\|\psi\|^2_{\mathcal{H}}$ par $g(\|\psi\|^2_{\mathcal{H}})$, où $g(\cdot)$ est une fonction monotone croissante sur $\mathbb{R}_+$. La fonctionnelle de risque régularisée s'écrit sous la forme

$$\frac{1}{n}\sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta\, g(\|\psi\|^2_{\mathcal{H}}). \tag{1.3}$$

Le choix du paramètre de régularisation η, appelé souvent hyperparamètre en apprentissage statistique, est crucial pour contrôler le compromis entre erreur d'apprentissage et degré de régularité. Il est donc lié directement à l'erreur de généralisation et conditionne la convergence de l'algorithme. Dans ce manuscrit, nous n'aborderons pas son optimalité en précisant que celle-ci est étudiée par Vapnik [Vap95], ainsi que par Wahba, Lin et Zhang [WLZ00] pour un cas particulier de fonctions coût, celui des SVM. Un exemple plus général a été considéré plus récemment par Cucker et Smale dans [CS02b], ou encore dans [CH02] et [Cap06].

Un nombre important de méthodes d'apprentissage a été développé dans ce cadre. Celles-ci se différencient principalement par deux caractéristiques clés : d'une part la fonctionnelle de coût $V(f(x_i), y_i)$ à minimiser, et d'autre part l'espace fonctionnel $\mathcal{H}$ des fonctions candidates. Cette thèse traite d'une catégorie d'espaces fonctionnels particuliers, les espaces de Hilbert à noyau reproduisant. Les méthodes associées constituent ce qu'on appelle couramment les méthodes à noyaux. Avant de continuer, il convient à présent de préciser qu'il existe une interprétation Bayésienne aux méthodes d'apprentissage statistique régularisées.

Sous un angle probabiliste

La communauté traitant de reconnaissance des formes peut être divisée en 2 groupes : les fréquentistes et les probabilistes (ou Bayésiens). Les premiers traitent les observations des phénomènes sans considérer les lois de probabilité les ayant engendrées. Les seconds cherchent à remonter aux distributions de probabilité à partir des observations par des techniques dites d'inférence. Bien que l'on adopte la première philosophie tout au long de ce manuscrit, il est toutefois intéressant de préciser qu'il existe un lien entre les deux approches, ou plus précisément une interprétation probabiliste des méthodes régularisées selon Tikhonov. On désigne par $P(\mathcal{A}_n|\psi)$ la probabilité conditionnelle d'avoir l'ensemble d'apprentissage $\mathcal{A}_n$ ayant la fonction ψ, et par $P(\psi|\mathcal{A}_n)$ la probabilité conditionnelle de ψ sachant $\mathcal{A}_n$. Par la règle de Bayes, la probabilité *a posteriori* $P(\psi|\mathcal{A}_n)$ est alors donnée par

$$P(\psi|\mathcal{A}_n) = P(\mathcal{A}_n|\psi)\, P(\psi), \tag{1.4}$$

où $P(\psi)$ désigne la probabilité *a priori* de la fonction ψ. On suppose dans la suite que les échantillons suivent le modèle $y_i = \psi(x_i) + e_i$, où le bruit e_i est distribué selon une loi normale de variance σ^2. On peut alors écrire la probabilité $P(\mathcal{A}_n|\psi)$ selon

$$P(\mathcal{A}_n|\psi) = \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \psi(x_i))^2 \Big).$$

La probabilité *a posteriori* est donnée par (1.4), avec

$$P(\psi|\mathcal{A}_n) = \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \psi(x_i))^2 \Big)\, P(\psi).$$

On a recours à la technique du maximum *a posteriori* pour estimer ψ^*, ce qui conduit à la minimisation de l'expression

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \psi(x_i))^2 + \frac{2\sigma^2}{n} \log(1/P(\psi)). \tag{1.5}$$

On retrouve les deux caractéristiques clés de l'apprentissage statistique : le premier terme de l'expression correspond au risque empirique de fonction coût quadratique $V(\psi(x_i), y_i) = (y_i - \psi(x_i))^2$. Le second terme est un terme de pénalisation. En supposant que la probabilité *a priori* puisse s'écrire sous la forme $P(\psi) = \exp(-g(\|\psi\|^2_{\mathcal{H}}))$, on retrouve la régularisation de Tikhonov dans (1.3), avec $2\sigma^2/n$ pour paramètre de régularisation. La fonction

ψ^* obtenue par la minimisation du critère (1.5) est optimale sous réserve que le bruit suive une loi normale. Pour une densité de probabilité Laplacienne, $P(e_i) = (1/2)\exp^{-|e_i|}$, la fonction coût correspondante est le coût L_1, à savoir $V(\psi(x_i), y_i) = |y_i - \psi(x_i)|$. L'optimalité de la régression logistique avec $V(\psi(x_i), y_i) = \log(1 + \exp(-y_i\psi(x_i)))$ est étudiée dans [Zha04], avec d'autres fonctions coûts, ou encore dans [SS01] avec la maximisation de la vraisemblance. Plus généralement, l'approche par maximisation de la probabilité *a posteriori* est étudiée par Mackay [Mac03] dans le cadre des processus Gaussiens.

1.2 Noyau reproduisant

Dans le paragraphe précédent, nous avons proposé de restreindre l'espace de recherche de ψ à un espace de Hilbert à noyau reproduisant. Depuis les travaux précurseurs de Aronszajn dans [Aro50] sur les noyaux reproduisants, on a de plus en plus recours à ce type d'espaces, notamment depuis qu'ils ont été retenus pour la résolution de problèmes d'interpolation par Parzen [Par70], Kailath [Kai71] et Wahba [Wah90]. Plus récemment, on peut se référer aux travaux de Saito [SABO99]. On présente ici un aperçu succinct des noyaux reproduisants avant d'exposer à la section suivante les éléments clés qui ont contribué à leurs succès dans le cadre de problèmes d'apprentissage. Dans ce qui suit, on considère un espace mesurable $\mathcal{X}$ auquel on associe le produit scalaire $\langle \cdot, \cdot \rangle$ et la norme correspondante $\| \cdot \|^2$.

1.2.1 Espace de Hilbert à noyau reproduisant

Un noyau désigne une fonction $\kappa(\cdot, \cdot)$ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$, à symétrie Hermitienne, c'est-à-dire telle que $\kappa(x_i, x_j) = \overline{\kappa}(x_j, x_i)$ pour tout $x_i, x_j \in \mathcal{X}$. On rappelle les deux définitions fondamentales suivantes et les propriétés qui en découlent, en renvoyant le lecteur vers [Aro50, CS02c] pour plus de détails.

Définition 1 (Noyau défini positif)**.** *Un noyau κ est dit défini positif sur $\mathcal{X}$ s'il vérifie*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \, \overline{a_j} \, \kappa(x_i, x_j) \geq 0 \tag{1.6}$$

pour tout $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ et $a_1, \ldots, a_n \in \mathbb{C}$.

Les noyaux définis positifs sont considérés comme une généralisation du produit scalaire. Ce dernier constitue en effet un noyau défini positif particulier, dont certaines des propriétés sont vérifiées par tout noyau défini positif. C'est

le cas par exemple de l'inégalité de Cauchy-Schwartz : pour tout noyau $\kappa(\cdot,\cdot)$ défini positif sur $\mathcal{X}$, et pour tout $x_i, x_j \in \mathcal{X}$, on a

$$|\kappa(x_i, x_j)|^2 \leq \kappa(x_i, x_i)\,\kappa(x_j, x_j).$$

Bien que la linéarité ne soit évidemment pas vérifiée par les noyaux définis positifs en général, cette généralisation s'avère très intéressante comme on le montre à la Section 1.3.1 au travers du coup du noyau.

Définition 2 (Noyau reproduisant et RKHS)**.** *Soit $(\mathcal{H}, \langle\cdot\,,\cdot\rangle_{\mathcal{H}})$ un espace de Hilbert constitué de fonctions de $\mathcal{X}$ dans $\mathbb{C}$. La fonction $\kappa(x_i, x_j)$ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$ est le noyau reproduisant de $\mathcal{H}$, sous réserve que celui-ci en admette un, si et seulement si*

- *la fonction $\kappa_{x_i} : x_j \mapsto \kappa_{x_i}(x_j) = \kappa(x_i, x_j)$ appartient à $\mathcal{H}$, quel que soit $x_i \in \mathcal{X}$ fixé ;*
- *on a $\psi(x_j) = \langle \psi, \kappa_{x_j}\rangle_{\mathcal{H}}$ pour tout $x_j \in \mathcal{X}$ et $\psi \in \mathcal{H}$.*

On dit que $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant, ou encore un RKHS, acronyme de Reproducing Kernel Hilbert Space.

D'après le premier point de la Définition 2, on désigne par κ_{x_i} et $\kappa(x_i, \cdot)$ la même fonction, qu'on appelle fonction noyau. Le second point de la Définition 2 est connu sous le nom de *propriété reproduisante*. Celle-ci, ainsi que l'existence du noyau reproduisant, sont liées directement au Théorème de Représentation de Riesz. En effet, soit une fonctionnelle $\psi(\cdot) \in \mathcal{H}$ et $\psi(x)$ son évaluation continue pour tout $x \in \mathcal{X}$. Selon ce théorème, il existe une fonction $h_x(\cdot) \in \mathcal{H}$ telle que $\psi(x) = \langle \psi, h_x\rangle_{\mathcal{H}}$. On retrouve la propriété reproduisante et, en posant $\kappa(x_i, x_j) = h_{x_i}(x_j)$ pour tout $x_i, x_j \in \mathcal{X}$, on retrouve le noyau reproduisant. On représente à la Figure 1.1 l'espace des données $\mathcal{X}$ et l'espace $\mathcal{H}$ induit par le noyau reproduisant κ.

Afin de fixer les idées, on présente une analogie noyau reproduisant et RKHS d'une part, et fonction indicatrice et espace $\mathcal{L}_2[a,b]$ d'autre part où toute fonction ψ vérifie $\|\psi\|_2^2 = \int_a^b |\psi(x)|^2\,\mathrm{d}x < \infty$. Pour tout $x \in [a,b]$, la fonction indicatrice $\mathbb{1}_x$ permet l'évaluation de toute fonction ψ en x, selon $\mathbb{1}_x : \psi \mapsto \mathbb{1}_x\psi = \psi(x)$. On retrouve ainsi la propriété du noyau reproduisant dans un RKHS. L'espace $\mathcal{L}_2[a,b]$ n'est toutefois pas un RKHS car que la fonction indicatrice $\mathbb{1}_x$ n'appartient pas à cet espace, contrairement à la fonction noyau κ_x qui est dans son RKHS.

Corollaire 3 (Coup du noyau)**.** *Tout noyau reproduisant κ d'un espace de Hilbert $\mathcal{H}$ s'écrit comme un produit scalaire dans cet espace, selon*

$$\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j}\rangle_{\mathcal{H}}$$

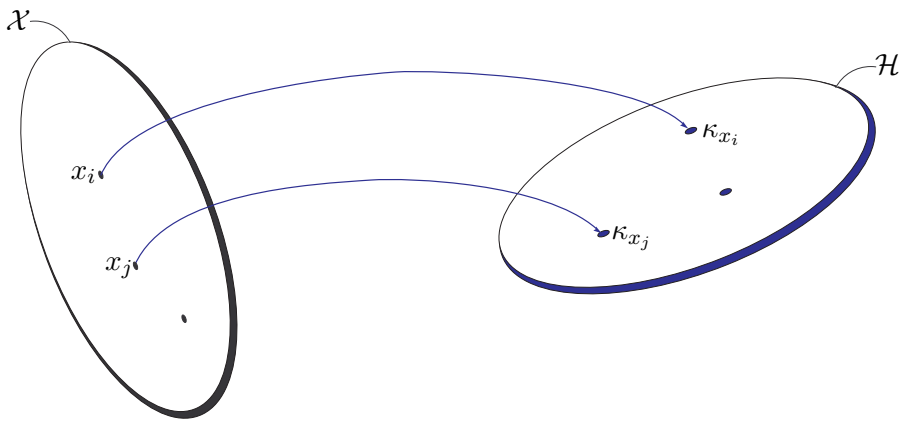


FIGURE 1.1 – Espace des données $\mathcal{X}$ et espace $\mathcal{H}$ induit par le noyau reproduisant κ.

pour tout $x_i, x_j \in \mathcal{X}$.

Ce corollaire constitue une propriété fondamentale des noyaux reproduisants et des méthodes à noyaux. Il sera étudié plus en détails dans la Section 1.3.1. Sa démonstration découle directement de la propriété reproduisante $\psi(x_j) = \langle \psi, \kappa_{x_j} \rangle_{\mathcal{H}}$ des noyaux. Il suffit pour cela de remplaçer ψ dans cette expression par la fonction noyau κ_{x_i}.

Théorème 4 (Moore-Aronszajn)**.** *A tout noyau défini positif, il correspond un RKHS unique, et réciproquement.*

Démonstration. On montre tout d'abord que tout noyau reproduisant est défini positif. Pour cela, il suffit de constater que $\sum_i \sum_j a_i \overline{a_j} \kappa(x_i, x_j) = \| \sum_i a_i \kappa_{x_i} \|^2$ ne peut être négatif. Réciproquement, on démontre que tout noyau défini positif κ est le noyau reproduisant d'un espace de Hilbert de fonctions de $\mathcal{X}$ dans $\mathbb{C}$. Pour cela, on considère l'espace vectoriel $\mathcal{H}'$ engendré par l'ensemble des fonctions $\{\kappa_{x_i}\}$ pour $x_i \in \mathcal{X}$. Ceci permet d'exprimer tout élément de $\mathcal{H}'$ comme une combinaison linéaire finie ($n < \infty$) de ces fonctions selon

$$\mathcal{H}' = \Big\{ \psi : \psi(\cdot) = \sum_{i=1}^{n} a_i \, \kappa_{x_i}(\cdot), \; x_i \in \mathcal{X}, \; a_i \in \mathbb{C} \Big\}.$$

A cet espace, on associe le produit scalaire

$$\langle \psi, \phi \rangle_{\mathcal{H}'} = \Big\langle \sum_{i=1}^{n} a_i \ \kappa_{x_i}, \sum_{j=1}^{n} \ b_j \ \kappa_{x_j} \Big\rangle,$$

avec $\psi = \sum_{i=1}^{n} a_i \ \kappa_{x_i}$ et $\phi = \sum_{j=1}^{n} b_j \ \kappa_{x_j}$ appartenant à l'espace $\mathcal{H}'$. En ré-arrangeant les sommations, et par le coup du noyau $\langle \kappa_{x_i}, \kappa_{x_j} \rangle = \kappa(x_i, x_j)$, on peut simplifier l'expression du produit scalaire en

$$\langle \psi, \phi \rangle_{\mathcal{H}'} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \ \overline{b_j} \ \kappa(x_i, x_j).$$

Muni de ce produit scalaire, l'espace ainsi construit est un espace pré-Hilbertien. Pour obtenir un espace de Hilbert, il suffit de le compléter conformément à [Aro50] de sorte que toute suite de Cauchy y converge. ■

Le Théorème de Moore-Aronszajn établit le lien entre noyau défini positif et RKHS. Par abus de langage, on remplace dans la suite la dénomination de noyau défini positif par noyau reproduisant. Avant de préciser leurs propriétés fondamentales dans le cadre de méthodes d'apprentissage régularisées, on présente les noyaux reproduisants les plus connus ainsi que certaines règles permettant de les combiner afin d'en obtenir de nouveaux.

1.2.2 Noyaux reproduisants : construction et exemples

On présente ici des techniques pour construire des noyaux reproduisants, et quelques exemples couramment utilisés. On renvoie le lecteur à [Vap95, Her02, STC04] pour d'autres propriétés des noyaux reproduisants, ainsi que diverses règles permettant de les combiner.

Théorème 5. *Soit κ_1 et κ_2 deux noyaux reproduisants de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$. La fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ est un noyau reproduisant s'il est défini par une de ces expressions, pour tout $x_i, x_j \in \mathcal{X}$,*

1. *combinaison linéaire : $\kappa(x_i, x_j) = \beta_1 \kappa_1(x_i, x_j) + \beta_2 \kappa_2(x_i, x_j)$, pour tout $\beta_1, \beta_2 \in \mathbb{R}_+$.*
2. *décalage : $\kappa(x_i, x_j) = \kappa_1(x_i, x_j) + c$, pour tout $c \in \mathbb{R}_+$.*
3. *produit : $\kappa(x_i, x_j) = \kappa_1(x_i, x_j) \, \kappa_2(x_i, x_j)$.*
4. *exposant : $\kappa(x_i, x_j) = \kappa_1(x_i, x_j)^p$, pour tout $p \in \mathbb{N}_+$.*
5. *exponentiel : $\kappa(x_i, x_j) = \exp(\kappa_1(x_i, x_j)/\sigma^2)$, pour tout $\sigma \in \mathbb{R}$.*

6. *normalisation :* $\kappa(x_i, x_j) = \frac{\kappa_1(x_i,x_j)}{\sqrt{\kappa_1(x_i,x_i)\,\kappa_1(x_j,x_j)}}$.

éléments de démonstration. A partir du Théorème 4 de Moore-Aronszajn, il suffit de démontrer pour chaque cas que le noyau κ est défini positif, à savoir $\sum_i \sum_j a_i\ \overline{a_j}\ \kappa(x_i, x_j) \geq 0$ pour tout $x_i, x_j \in \mathcal{X}$ et $a_i, a_j \in \mathbb{C}$. On peut facilement montrer cela pour les quatre premiers points. Pour le cinquième point, il suffit de décomposer l'exponentielle en un développement de noyau κ_1 de puissances différentes. Finalement, pour le noyau normalisé, il suffit de remarquer que le dénominateur est toujours positif, et que le numérateur est défini positif. ■

Les noyaux reproduisants classiques sont principalement regroupés en deux catégories : les noyaux radiaux et les noyaux projectifs. Ces derniers dépendent du produit scalaire, comme le noyau linéaire $\kappa(x_i, x_j) = \langle x_i, x_j \rangle$ ou encore les noyaux polynômiaux $\kappa(x_i, x_j) = \langle x_i, x_j \rangle^p$ de degré $p \in \mathbb{N}^*$. La règle du décalage (règle 2) conduit au noyau polynomial complet $\kappa(x_i, x_j) = (\langle x_i, x_j \rangle + c)^p$. Le noyau exponentiel, soit $\kappa(x_i, x_j) = \exp(\langle x_i, x_j \rangle / \sigma_0^2)$ où σ_0 est un paramètre fixe correspondant à la largeur du noyau, résulte de l'application de la règle 5 au noyau linéaire. Le plus connu des noyaux radiaux est sans doute le noyau Gaussien, défini par $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \sigma_0^2)$, que l'on obtient en normalisant (règle 6) le noyau exponentiel. Le noyau de Laplace, $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\| / \sigma_0)$, fait aussi parti de cette catégorie. D'autres noyaux reproduisants sont obtenus en combinant des noyaux classiques selon le Théorème 5. On présente en Annexe A les noyaux reproduisants les plus utilisés dans le cadre des méthodes de reconnaissance des formes. Toutefois, cette liste n'est pas exhaustive puisqu'en les combinant entre eux, on en obtient de nouveaux sous réserve de conserver le caractère défini positif.

1.3 Au-delà du modèle linéaire : méthodes à noyaux

Les méthodes à noyaux sont pour la plupart issues d'algorithmes linéaires auxquels on a pu appliquer les deux résultats clés que sont le coup du noyau [ABR64] et le Théorème de Représentation [Wah90, SHW00]. Dans ce qui suit, on note $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un ensemble d'apprentissage donné avec $x_i \in \mathcal{X}$ les données d'entrée et $y_i \in \mathbb{R}$ les sorties désirées.

1.3.1 Coup du noyau

On rappelle la propriété fondamentale introduite au Corollaire 3. Tout noyau reproduisant s'écrit

$$\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}}$$

quels que soient $x_i, x_j \in \mathcal{X}$, avec $\mathcal{H}$ l'espace de Hilbert induit par ce noyau. Ainsi, le noyau $\kappa(x_i, x_j)$ fournit le produit scalaire dans $\mathcal{H}$ des images κ_{x_i} et κ_{x_j} de toute paire d'éléments x_i et x_j de $\mathcal{X}$, sans qu'il soit nécessaire d'expliciter ces images. Ce principe, connu sous le nom du *coup du noyau* ou *kernel trick* en anglais, permet de transformer les méthodes linéaires de traitement de données en des méthodes non-linéaires, sous réserve qu'elles puissent s'exprimer uniquement en fonction de produits scalaires des observations. Pour cela, il suffit de remplacer chacun de ces produits scalaires $\langle x_i, x_j \rangle$, qui n'est autre que le noyau linéaire, par un noyau non-linéaire $\kappa(x_i, x_j)$. Ainsi la structure des algorithmes demeure-t-elle inchangée, et le surcoût calculatoire dû à l'évaluation des noyaux négligeable.

On souligne l'importance du noyau pour le calcul implicite d'un produit scalaire dans l'espace $\mathcal{H}$. Cet espace est très souvent de dimension $d_{\mathcal{H}}$ supérieure à celle de l'espace des observations $d_{\mathcal{X}}$. En utilisant par exemple un noyau Gaussien $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma_0^2)$, l'espace induit est de dimension infinie. Sans la mise en œuvre du coup du noyau, la détermination du produit scalaire dans de tels espaces serait impossible.

Géométrie associée

L'intérêt du coup du noyau va au-delà d'une simple évaluation de produit scalaire, en incluant l'évaluation de distances et d'angles dans l'espace RKHS. La distance dans $\mathcal{H}$ entre 2 fonctions κ_{x_i} et κ_{x_j} s'exprime par $\|\kappa_{x_i} - \kappa_{x_j}\|^2_{\mathcal{H}} = \langle \kappa_{x_i} - \kappa_{x_j}, \kappa_{x_i} - \kappa_{x_j} \rangle_{\mathcal{H}}$. En développant cette expression, on obtient

$$\|\kappa_{x_i} - \kappa_{x_j}\|^2_{\mathcal{H}} = \kappa(x_i, x_i) - 2\text{Ré}\{\kappa(x_i, x_j)\} + \kappa(x_j, x_j), \tag{1.7}$$

où Ré$(\cdot)$ désigne la partie réelle. Ainsi peut-on calculer la distance entre les éléments de cet espace sans qu'il soit nécessaire de les expliciter. On retrouve l'esprit du coup du noyau, qui permet selon cette expression de transformer toute méthode linéaire qui ne dépendrait uniquement que des distances entre les différents éléments en une méthode non-linéaire basée sur les noyaux. Ceci est le cas de la méthode des k-plus-proches-voisins par exemple. Puisque

pour tout $x \in \mathcal{X}$ on a $\kappa(x,x) = \langle \kappa_x, \kappa_x \rangle_{\mathcal{H}} = \|\kappa_x\|^2_{\mathcal{H}}$, on peut aussi réécrire l'équation (1.7) pour un noyau à valeurs réelles selon

$$\kappa(x_i, x_j) = \frac{1}{2}\left(\|\kappa_{x_i}\|^2_{\mathcal{H}} + \|\kappa_{x_j}\|^2_{\mathcal{H}} - \|\kappa_{x_i} - \kappa_{x_j}\|^2_{\mathcal{H}}\right).$$

Cette équation montre que le noyau $\kappa(x_i, x_j)$ est une mesure de similitude entre x_i et x_j, celle-ci étant l'opposé du carré de la distance entre leurs images dans l'espace fonctionnel à deux termes additifs près. Pour le calcul de l'angle entre deux fonctions noyau κ_{x_i} et κ_{x_j}, il suffit de réécrire le cosinus de l'angle en termes de noyau selon

$$\cos(\kappa_{x_i}, \kappa_{x_j}) = \frac{\langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}}}{\|\kappa_{x_i}\|_{\mathcal{H}} \|\kappa_{x_j}\|_{\mathcal{H}}} = \frac{\kappa(x_i, x_j)}{\sqrt{\kappa(x_i, x_i)\kappa(x_j, x_j)}}.$$

Une fois de plus, le coup du noyau permet d'exprimer une mesure dans $\mathcal{H}$ sans qu'il soit nécessaire d'exhiber les éléments de cet espace.

Transformation unitaire

Bien que l'espace de Hilbert $\mathcal{H}$ induit par un noyau reproduisant donné κ soit unique, rien ne s'oppose à l'élaboration d'un tout autre espace muni du même produit scalaire. En effet, pour tout opérateur unitaire $\mathbf{U}$ de $\mathcal{H}$ vers $\mathcal{H}_{\mathbf{U}}$, le produit scalaire dans ce dernier est

$$\langle \mathbf{U}\kappa_{x_i}, \mathbf{U}\kappa_{x_j} \rangle_{\mathcal{H}_{\mathbf{U}}} = \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}},$$

qui n'est autre que le noyau reproduisant $\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}}$. On dit alors que les espaces $\mathcal{H}$ et $\mathcal{H}_{\mathbf{U}}$ sont isomorphes, ou encore que le RKHS $\mathcal{H}$ est unique à un isomorphisme près. Pour plus de détails sur l'équivalence des espaces induits par le même noyau, on renvoie le lecteur intéressé vers [Ste02, MNY06].

1.3.2 Théorème de Représentation

Le coup du noyau offre une interprétation du noyau reproduisant en tant que produit scalaire, et permet d'élaborer des méthodes non-linéaires à partir d'algorithmes linéaires. Pour que ce principe soit opérationnel, il nécessite souvent d'être associé au Théorème de Représentation. Ce dernier, à usage multidisciplinaire aujourd'hui, est issu des travaux précurseurs de Kimeldorf et Wahba dans le domaine de la théorie de l'approximation [KW71, Wah90]. Plus récemment, il a été repris dans le cadre de la résolution de problèmes

inverses [Kur04], ainsi qu'en théorie de l'apprentissage [CS02c]. La formulation suivante du Théorème de Représentation et sa démonstration pour différents types de fonctions coûts sont principalement dues à Schölkopf *et coll.* [SHW00].

Théorème 6. *Soient $\mathcal{X}$ un compact, $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un ensemble d'apprentissage donné avec $x_i \in \mathcal{X}$ l'ensemble des données et $y_i \in \mathbb{C}$ l'ensemble des sorties désirées, V une fonction coût arbitraire et $g(\cdot)$ une fonction monotone croissante sur $\mathbb{R}+$. Soit $\mathcal{H}$ un espace de Hilbert induit par le noyau κ défini positif sur $\mathcal{X}$. Toute fonction $\psi^* \in \mathcal{H}$ minimisant la fonctionnelle de risque régularisée*

$$\frac{1}{n}\sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta\, g(\|\psi\|_{\mathcal{H}}^2), \tag{1.8}$$

peut s'écrire sous la forme

$$\psi^*(\cdot) = \sum_{j=1}^{n} \alpha_j^*\ \kappa(x_j, \cdot). \tag{1.9}$$

Démonstration. Soit $\mathcal{H}_n$ le sous-espace de $\mathcal{H}$ engendré par les fonctions $\{\kappa(x_1, \cdot), \ldots, \kappa(x_n, \cdot)\}$, c'est-à-dire

$$\mathcal{H}_n = \Big\{\psi \in \mathcal{H} : \psi(\cdot) = \sum_{j=1}^{n} \alpha_j\ \kappa(x_j, \cdot),\ \alpha_1, \ldots, \alpha_n \in \mathbb{C}\Big\}.$$

Toute fonction ψ de $\mathcal{H}$ admet une et une seule décomposition en deux contributions, l'une appartenant à l'espace $\mathcal{H}_n$ et l'autre qui lui est orthogonale. On peut en effet écrire

$$\psi = \psi^* + \psi^\perp,$$

avec $\psi^* \in \mathcal{H}_n$ et $\psi^\perp$ la composante orthogonale telles que $\langle \psi^\perp, \kappa_{x_i}\rangle_{\mathcal{H}} = 0$ pour tout $i = 1, \ldots, n$. La propriété reproduisante permet d'évaluer ψ en x_i selon l'expression

$$\begin{aligned} \psi(x_i) &= \langle \psi, \kappa_{x_i}\rangle_{\mathcal{H}} \\ &= \sum_{j=1}^{n} \alpha_j \langle \kappa_{x_j}, \kappa_{x_i}\rangle_{\mathcal{H}} + \langle \psi^\perp, \kappa_{x_i}\rangle_{\mathcal{H}}. \end{aligned}$$

Puisque le second terme s'annule par orthogonalité, on obtient l'expression

$$\psi(x_i) = \sum_{j=1}^{n} \alpha_j \kappa(x_i, x_j).$$

Comme les évaluations de ψ en chaque point de l'ensemble d'apprentissage ne dépendent que des coefficients $\{\alpha_1, \ldots, \alpha_n\}$, le risque empirique (1.8) ne dépend pas de la composante orthogonale $\psi^\perp$. En minimisant ce risque, on obtient la classe des fonctions équivalentes dans $\mathcal{H}$ telle que deux fonctions ψ et ϕ appartiennent à la même classe si et seulement si $\psi(x_i) = \phi(x_i)$ pour tout $i = 1, \ldots, n$. Il reste à déterminer $\psi^\perp$ pour une classe de fonctions équivalentes donnée afin de minimiser le terme régularisant. Par le théorème de Pythagore dans $\mathcal{H}$ appliqué à ψ, soit $\|\psi\|^2_{\mathcal{H}} = \|\psi^*\|^2_{\mathcal{H}} + \|\psi^\perp\|^2_{\mathcal{H}}$, le terme régularisant $g(\|\psi\|^2_{\mathcal{H}})$ dans (1.8) s'écrit

$$g(\|\psi\|^2_{\mathcal{H}}) = g\Big(\big\|\sum_{j=1}^{n} \alpha_j \ \kappa_{x_j}\big\|^2_{\mathcal{H}} + \big\|\psi^\perp\big\|^2_{\mathcal{H}}\Big).$$

Comme $g(\cdot)$ est monotone croissante, la fonction qui minimise l'expression ci-dessus, pour une classe de fonctions équivalentes donnée, doit vérifier $\|\psi^\perp\|^2_{\mathcal{H}} = 0$. ■

L'importance de ce théorème réside dans l'existence d'une solution unique à une fonctionnelle de coût régularisée, celle-ci pouvant s'exprimer comme un développement en série fini de fonctions noyau. La minimisation de cette fonction coût (1.8) se ramène à un problème d'optimisation à n dimensions, celui de la détermination des coefficients optimaux $\alpha_1^*, \ldots, \alpha_n^* \in \mathbb{C}$.

1.4 Exemples de méthodes à noyaux

Pour fixer les idées, nous introduisons ici quelques exemples de méthodes d'apprentissage statistiques, les *Support Vector Machines* faisant l'objet du chapitre suivant. Pour ces méthodes, la fonctionnelle de risque à minimiser est pénalisée selon le principe de Tikhonov, soit

$$\frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta\, g(\|\psi\|^2_{\mathcal{H}}),$$

où $\mathcal{H}$ désigne un RKHS de noyau reproduisant donné. Comme évoqué précédemment, cette régularisation pénalise les variations importantes de la

fonction ψ. Pour s'en convaincre, il suffit de borner l'évaluation de ψ en tout point $x \in \mathcal{X}$ selon

$$|\psi(x)|^2 = |\langle \psi, \kappa_x \rangle|^2 \leq \|\psi\|^2 \ \|\kappa_x\|^2 = \|\psi\|^2 \ \kappa(x,x), \tag{1.10}$$

à partir de la propriété reproduisante et de l'inégalité de Cauchy-Schwartz. En majorant la norme dans un RKHS par $\|\psi\| \leq \tau$, on obtient une borne supérieure pour les valeurs de cette fonction en tout point de $\mathcal{X}$, avec $|\psi(x)| \leq M\tau$ où $M = \sqrt{\kappa(x,x)}$.

Pour une analyse non-supervisée, on cherche à extraire la structure sous-jacente des observations, l'ensemble d'apprentissage étant constitué de données non-étiquetées $\mathcal{A}_n = \{x_1, \ldots, x_n\}$. Parmi les méthodes existantes, l'analyse en composantes principales est incontestablement la plus connue. Celle-ci détermine un jeu d'axes orthogonaux tel que la variance de l'ensemble d'apprentissage projeté sur celui-ci est maximum. La formulation classique du problème consiste à atténuer la contrainte de régularisation dans l'expression (1.3), en imposant une norme unité aux axes recherchés, similaire à la régularisation de type Ivanov [Iva76]. Le premier axe principal s'obtient alors par maximisation de la fonction coût $|\psi(x_i)|^2$ où $\psi(x_i) = \langle \psi, \kappa_{x_i} \rangle_{\mathcal{H}}$ représente la projection de κ_{x_i} sur l'axe principal défini par ψ dans $\mathcal{H}$. Le risque empirique à minimiser est alors donné par

$$-\frac{1}{n} \sum_{i=1}^{n} |\psi(x_i)|^2,$$

sous la contrainte de normalisation $\|\psi\|_{\mathcal{H}}^2 = 1$. Les axes principaux suivants sont construits à partir de la même fonction coût tout en étant orthogonaux au premier, et entre eux. Dans cette formulation, on a supposé que les fonctions noyau κ_{x_i} sont centrées dans l'espace $\mathcal{H}$. Dans le cas contraire, la fonction coût s'écrit selon $|\psi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \psi(x_j)|^2$.

Dans le cas d'un problème de discrimination, l'ensemble d'apprentissage est formé des données x_i et de leurs étiquettes y_i. La fonction ψ^* recherchée vise à permettre l'identification de l'étiquette à partir d'une donnée. Ceci doit alors être vrai ou presque pour tout couple (x_i, y_i) de l'ensemble d'apprentissage, ce qui correspond à la contrainte $\psi^*(x_i) = y_i$. Pour un problème de classification à 2 classes, les étiquettes peuvent être codées selon $y_i \in \{\pm 1\}$, ce qui se traduit alors par la contrainte $y_i \psi^*(x_i) = 1$. Ainsi cherche-t-on à minimiser la fonctionnelle risque suivante

$$\frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \psi(x_i)).$$

On retrouve le coût de la marge diffuse, connue par coût charnière ou *soft margin* en anglais, qui a contribué au succès des Support Vector Machines (SVM). Le terme de régularisation s'écrit sous la forme $\eta\ g(\|\psi\|^2_{\mathcal{H}})$ où le paramètre de pénalisation η contrôle le compromis entre la complexité de la solution et l'erreur d'apprentissage.

Plus généralement, on peut s'intéresser au problème de moindres carrés correspondant à la minimisation du risque empirique

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \psi(x_i))^2.$$

Par l'usage de la pénalisation quadratique $\|\psi\|^2_{\mathcal{H}}$, la solution de ce problème est obtenue par la résolution d'un système linéaire. Différentes techniques à complexité linéaire peuvent être considérées pour le résoudre, comme dans [Rif02] où une technique de gradient conjugué est proposée. Contrairement au cas des SVM esquissé ci-dessus, on montre que la solution dépend de tous les éléments de l'ensemble d'apprentissage, les coefficients α_i du développement en noyau issu du Théorème de Représentation n'ayant aucune raison d'être nuls. Afin d'être en mesure de répondre à des problèmes d'apprentissage en-ligne où l'on est confronté à un flux de données, il est nécessaire de se préoccuper de la parcimonie de la solution.

Chapitre 2

Eléments de la théorie de l'apprentissage statistique et *Support Vector Machines*

Le domaine de la reconnaissance des formes connaît une révolution depuis le milieu des années 90 avec la théorie de l'apprentissage statistique et l'avènement des *Support Vector Machines* (SVM) pour la résolution de problèmes de détection, de classification et de régression. Une présentation synthétique de ces éléments est donnée dans ce chapitre. Il est organisé ainsi. On commence par un rappel sur les fondements de la théorie de l'apprentissage statistique, plus particulièrement la VC-dimension et les principes de minimisation des risques empirique et structurel. Puis on présente l'algorithme classique des SVM dans les cas de données linéairement et non-linéairement séparables.

2.1 Introduction

Pour un problème d'apprentissage donné, les meilleures performances en généralisation peuvent être atteintes lorsqu'on trouve un compromis satisfaisant entre les performances atteintes sur l'ensemble d'apprentissage et la capacité d'apprentissage de la famille de statistiques considérée. Ce concept s'est concrétisé par la théorie de l'apprentissage statistique, élaborée au milieu des années 90 par Vapnik [Vap95] mais dont les premiers éléments remontent aux années 70 [VC71]. Les deux piliers de cette théorie sont les principes inductifs de la minimisation du risque empirique et du risque structurel. Cette théorie a mis en évidence une nouvelle classe de méthodes d'ap-

prentissage pour la reconnaissance des formes, les *Support Vector Machines* [BGV92]. Ces structures, couramment appelées SVM, constituent des solutions à marge maximale entre l'hyperplan séparateur qu'elles définissent et les échantillons d'un ensemble d'apprentissage. Des considérations fondamentales sur les espaces de Hilbert à noyau reproduisant, regroupées sous le principe du coup du noyau, leur assure une extension non-linéaire.

On présente dans ce chapitre l'algorithme classique de SVM, comme initialement proposé dans[BGV92] pour un problème de classification à deux classes. Toutefois, il existe une grande variété d'algorithmes selon les applications envisagées. Parmi ceux-ci, on trouve l'apprentissage à une classe, avec *one class SVM*, et les ν-SVM. Différents algorithmes de SVM pour les problèmes multi-classes ont été proposés tout au long de la dernière décennie, dont des techniques de types un-contre-tous [Vap95] (et amélioré dans [Vap98]) et un-contre-un [Kre99] ; Voir [Abe03] pour une comparaison des récentes techniques proposées. On peut aussi citer les travaux de Crammer *et coll.* qui proposent un cadre général de codage permettant d'adapter les algorithmes classiques de SVM au cas multi-classes [CS00, CS02a]. A l'exception de [EPM00], peu de travaux ont considéré les principes inductifs de minimisation des risque empirique et structurel afin d'élaborer un algorithme de SVM multi-classes approprié.

2.2 Eléments de théorie de l'apprentissage statistique

On reprend la théorie de l'apprentissage statistique, en introduisant des résultats clés sur les capacités et les performances en généralisation des méthodes d'apprentissage. Les éléments de cette section sont principalement extraits de [Vap95, Vap98].

2.2.1 Position du problème

On rappelle le schéma de l'apprentissage statistique dans le cadre d'un problème supervisé. On cherche à déterminer la relation qui lie un espace d'entrée $\mathcal{X}$ à un espace de sortie $\mathcal{Y}$. Cette relation, définie par un modèle ψ, minimise le risque réel

$$R(\psi) = \int_{\mathcal{X}\times\mathcal{Y}} V(\psi(x), y)\, \mathrm{d}P(x, y), \tag{2.1}$$

avec $\mathrm{d}P(x,y) = P(x,y)\,\mathrm{d}x\,\mathrm{d}y$, et V la fonctionnelle coût qui vise à pénaliser l'erreur entre la sortie estimée $\psi(x)$ et celle désirée y. On désigne par ψ^* la fonction minimisant le risque réel (2.1). La résolution de ce problème nécessite la connaissance de la distribution de probabilité conjointe $P(x,y)$, supposée inaccessible. On dispose au lieu de cela d'un ensemble de n couples échantillonnés aléatoirement selon cette distribution, que l'on désigne par $\mathcal{A}_n = \{(x_1,y_1),\ldots,(x_n,y_n)\}$, où $(x_i,y_i) \in \mathcal{X}\times\mathcal{Y}$. A partir de ces données, on cherche à minimiser le risque empirique

$$R_n(\psi) = \frac{1}{n}\sum_{i=1}^{n} V(\psi(x_i),y_i). \tag{2.2}$$

La minimisation du risque empirique est un problème mal-posé, puisqu'il existe une infinité de fonctions faisant état d'un apprentissage exacte. Toutefois, la plupart de ces fonctions ne déterminent pas les bonnes étiquettes sur de nouvelles données, pourtant obtenues à partir de la même distribution de probabilité P. Pour remédier à cet inconvénient, on a recours à une restriction du domaine des fonctions candidates. On considère par exemple des fonctions régulières reflétant la régularité du problème traité. Parmi un ensemble $\mathcal{F}$ de fonctions, on détermine alors la fonction ψ_n^* minimisant le risque empirique (2.2), selon

$$\psi_n^* = \arg\min_{\psi\in\mathcal{F}} R_n(\psi).$$

Rien ne garantit que la solution ainsi obtenue, au travers du processus d'apprentissage, ne soit optimale au sens de la minimisation du risque réel (2.1) dans la famille $\mathcal{F}$ des fonctions candidates. L'optimalité dans ce dernier cas peut s'exprimer selon l'expression

$$\psi_{\mathcal{F}}^* = \arg\min_{\psi\in\mathcal{F}} R(\psi).$$

Les fonctions ψ_n^*, $\psi_{\mathcal{F}}^*$ et ψ^* correspondent respectivement à la fonction obtenue par un processus d'apprentissage au sein de la famille $\mathcal{F}$, à celle minimisant le risque réel dans $\mathcal{F}$, et à la fonction optimale au sens de Bayes. Ces fonctions sont comparées entre elles à partir des risques en question. En considérant les deux premières, on désigne par erreur d'estimation, notée E_{estim}, la différence des risques réel et empirique au sein de $\mathcal{F}$, selon

$$E_{\text{estim.}} = R_n(\psi_n^*) - \inf_{\psi\in\mathcal{F}} R(\psi).$$

Cette erreur dépend de l'ensemble d'apprentissage et du processus d'apprentissage, ainsi que de $\mathcal{F}$. D'autre part, en considérant les deux dernières, on obtient l'erreur d'approximation, $E_{\text{approx.}}$, définie par

$$E_{\text{approx.}} = \inf_{\psi \in \mathcal{F}} R(\psi) - R(\psi^*).$$

Ne dépendant que du choix de $\mathcal{F}$, elle détermine la difficulté d'approximer le modèle optimal ψ^* à partir de cet ensemble de fonctions. En combinant les deux erreurs, l'erreur d'estimation et l'erreur d'approximation, on obtient l'erreur dite de modélisation définie selon l'expression

$$E_{\text{modél.}} = E_{\text{estim.}} + E_{\text{approx.}} = R_n(\psi_n^*) - R(\psi^*).$$

On illustre à la Figure 2.1 les différentes sources d'erreur induites par une procédure d'apprentissage visant à minimiser le risque empirique sur un ensemble d'apprentissage, en considérant une famille $\mathcal{F}$ de fonctions candidates. La minimisation de l'erreur de modélisation fait intervenir deux termes antagonistes, $E_{\text{estim.}}$ et $E_{\text{approx.}}$. Afin de diminuer la première, on a recours à une famille $\mathcal{F}$ plus riche en augmentant le nombre de fonctions candidates, au détriment de la seconde. Réciproquement, l'usage d'une famille $\mathcal{F}$ plus réduite entraîne une diminution de $E_{\text{estim.}}$ tandis que $E_{\text{approx.}}$ croît. Ce phénomène n'est autre que le fameux compromis biais-variance.

établie par les travaux précurseurs de Vapnik et Chervonenkis [VC71], la théorie de l'apprentissage statistique visent à traiter ce compromis en considérant deux problèmes. Dans un premier temps, la famille de modèles étant fixée, l'étude concerne la convergence uniforme du risque empirique vers le risque le plus proche du risque réel pour ces fonctions candidates. Il s'agit du principe inductif de minimisation du risque empirique. Dans un second temps, la question du choix de la famille de modèles est alors posée. La sélection de la famille optimale est alors considérée en recherchant le meilleur compromis possible. Il s'agit du principe inductif de minimisation du risque structurel.

2.2.2 Dimension de Vapnik-Chervonenkis

La dimension de Vapnik-Chervonenkis, ou encore VC-dimension, d'une famille de fonctions est une mesure de sa capacité d'apprentissage. Soit $\mathcal{F}$ une famille de fonctions définies sur $\mathcal{X}$.

Définition 7 (VC-dimension)**.** *La dimension h de Vapnik-Chervonenkis d'une famille $\mathcal{F}$ donnée est le plus grand nombre d'éléments de l'ensemble*

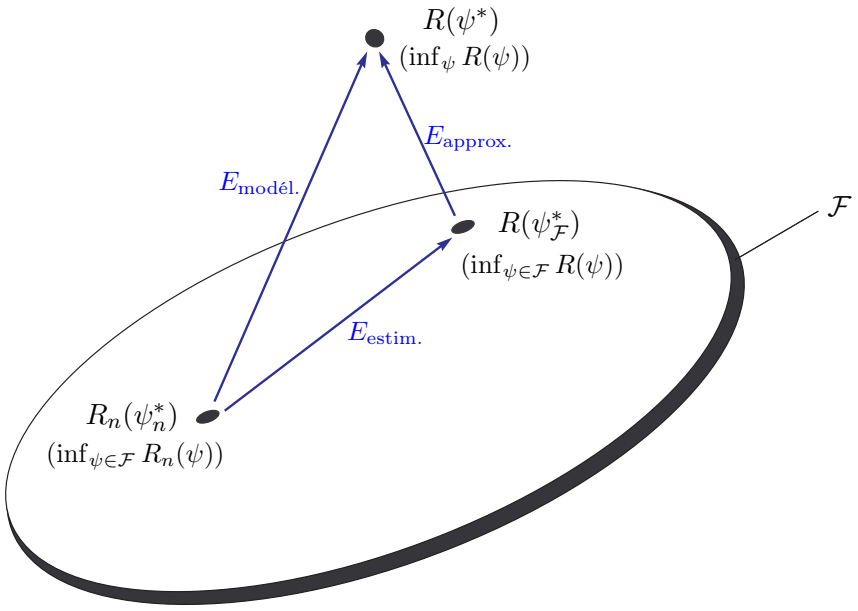


FIGURE 2.1 – Représentation schématique des différentes erreurs dues à la minimisation du risque empirique sur un ensemble $\mathcal{F}$ de fonctions.

des réalisations $\mathcal{X}$ dont les fonctions de $\mathcal{F}$ peuvent réaliser toutes les 2^h dichotomies.

Le sens d'une dichotomie est naturel pour une famille de fonctions à valeur binaire, soit $\{0,1\}$ dans le cas d'une famille de détecteurs, ou encore $\{-1,+1\}$ pour une classification. Cette dichotomie s'étend aisément aux familles de fonctions à valeurs réelles, au sens suivant : pour toute partition de l'espace des réalisations, $\mathcal{X}_+ \cup \mathcal{X}_- = \mathcal{X}$, il existe au moins une fonction ψ de $\mathcal{F}$ telle que l'hypothèse $\psi(x) \gtrless_{\mathcal{X}_-}^{\mathcal{X}_+} 0$ est vérifiée pour tout $x \in \mathcal{X}$. Il faut toutefois noter que cette expression est souvent considérée en remplaçant le seuil 0 par un seuil arbitraire.

Bien avant les développements proposés par Vapnik et Chervonenkis [VC71], l'intérêt d'une caractérisation de la capacité d'apprentissage d'une famille de classifieurs a été souligné par Cover dans le cadre des discriminants linéaires [Cov65]. Afin d'illustrer la notion de VC-dimension, on

s'intéresse précisément à présent à celle des discriminants linéaires dans l'espace $\mathbb{R}^l$, en commençant par $l = 1$ et $l = 2$. Dans $\mathbb{R}$, toute fonction appartenant à $\mathcal{F}$ est de la forme $\psi(x) = wx + w_0$. Comme illustré à la Figure 2.2 (a), deux points $\{x_1, x_2\}$ peuvent être discriminés par des fonctions de $\mathcal{F}$, en réalisant toutes les dichotomies que l'on représente sous forme de couples $\{\}/\{x_1, x_2\}, \{x_1\}/\{x_2\}, \{x_2\}/\{x_1\}, \{x_1, x_2\}/\{\}$. Cependant, pour trois points $\{x_1, x_2, x_3\}$ on ne peut pas obtenir les 2^3 dichotomies comme le montre la Figure 2.2 (b) avec le couple $\{x_1, x_3\}/\{x_2\}$. On dit alors que les discriminants linéaires de $\mathbb{R}$ admettent une VC-dimension égale à 2. En considérant $\mathbb{R}^2$, on peut avoir que toutes les dichotomies de 3 points peuvent être réalisées comme cela est présenté en Figure 2.2 (c). Il n'en est en revanche pas de même pour 4 points, la dichotomie $\{x_1, x_4\}/\{x_2, x_3\}$ étant impossible à obtenir comme le montre la Figure 2.2 (d). La VC-dimension des discriminants linéaires de $\mathbb{R}^2$ est donc égale à 3. On peut étendre aisément ce résultat à $\mathbb{R}^l$, où l'ensemble des discriminants linéaires $\psi(x) = \langle \boldsymbol{w}, x \rangle_{\mathcal{X}} + w_0$ admet une VC-dimension égale à $l+1$. Bien que la VC-dimension des discriminants linéaires corresponde au nombre de paramètres libres, cette analogie n'est pas vraie dans le cas général comme illustré par plusieurs contre-exemples dans [Vap95].

2.2.3 Principe de minimisation du risque empirique

La pertinence de la substitution du risque réel par le risque empirique est donné par le principe de minimisation du risque empirique. Le théorème suivant est un résultat clé de la théorie de l'apprentissage.

Théorème 8. *Soit $\mathcal{F}$ une famille de fonctions ψ. Le principe de minimisation du risque empirique est consistant si et seulement si le risque empirique converge uniformément vers le risque réel au sens probabiliste suivant :*

$$P(\sup_{\psi \in \mathcal{F}} \{R(\psi) - R_n(\psi)\} > \epsilon) \underset{n \to \infty}{\longrightarrow} 0$$

pour tout $\epsilon > 0$.

En pratique, on ne peut pas invoquer ce théorème puisqu'il fait appel à la distribution de probabilité inconnue $P(x, y)$ dans le calcul du risque réel R. La théorie de l'apprentissage statistique offre non seulement un résultat qualitatif sur la convergence uniforme du risque empirique vers le risque réel pour tout un ensemble des fonctions candidates, mais aussi des bornes explicites sur la vitesse de convergence de l'expression du Théorème 8. De plus, on montre qu'il existe une borne dite universelle ne dépendant pas

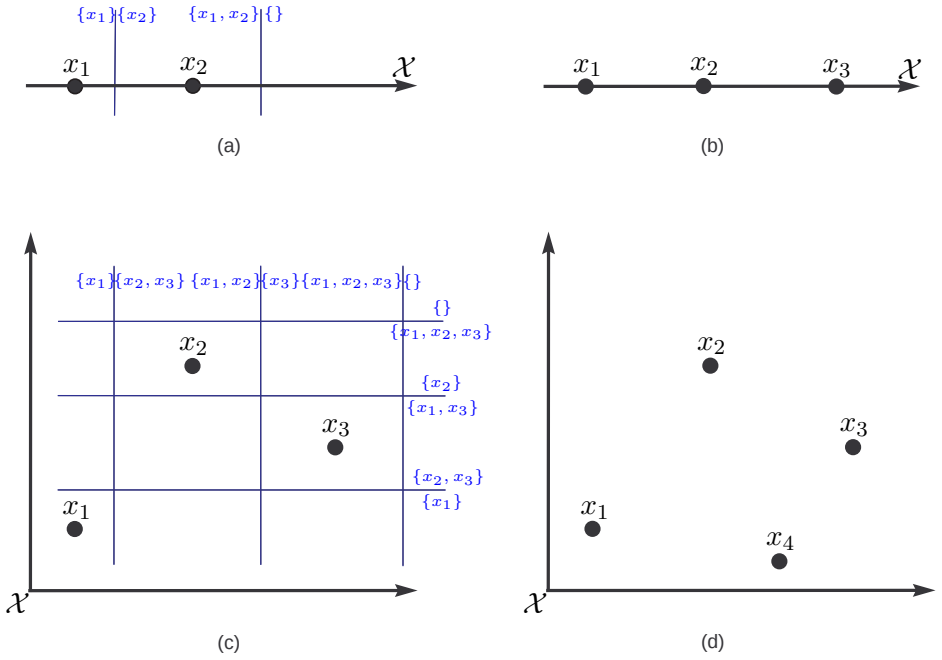


FIGURE 2.2 – Détermination de la VC-dimension des discriminants linéaires, de $\mathbb{R}$ (première ligne) et $\mathbb{R}^2$ (deuxième ligne).

de la distribution de probabilité des données. Plus encore, il est possible de démontrer que la borne suivante est satisfaite pour toute famille $\mathcal{F}$ de fonctions à VC-dimension h :

$$P(\sup_{\psi \in \mathcal{F}} \{R(\psi) - R_n(\psi)\} > \epsilon) \leq 4\,(2\,en/h)^h\ \mathrm{e}^{-n\epsilon^2/8}, \tag{2.3}$$

pour tout $\epsilon > 0$. Ce résultat permet de préciser le principe de minimisation du risque empirique présenté au Théorème 8. Avec une VC-dimension finie, la convergence uniforme est alors satisfaite, et le principe de minimisation du risque empirique est consistant. Il s'avère que cette condition est également nécessaire. Ce résultat, valable pour toute famille de fonctions, constitue une généralisation des bornes proposées séparément par Kolmogorov et Smirnov, largement répandues en statistique classique. Ces bornes classiques sont de plus valables asymptotiquement, comparées à l'expression (2.3) qui est satisfaite pour un nombre fini d'observations. Vapnik énonce dans [Vap95] d'autres inégalités, comme par exemple une borne supérieure sur l'écart entre les risques empirique et réel. Depuis sa formulation originelle,

plusieurs travaux ont revisité l'expression (2.3), en proposant des bornes de plus en plus fines. On pourra se référer à [Vay00] pour un aperçu sur ces différentes améliorations, dont

$$P(\sup_{\psi\in\mathcal{F}}\{R(\psi)-R_n(\psi)\}>\epsilon)\leq 4\,(2\,\mathrm{e}n/h)^h\;\mathrm{e}^{-n\epsilon^2}. \tag{2.4}$$

A partir de l'expression (2.4), il est possible de déduire un intervalle de confiance liant les risques empirique et réel pour un fonction appartenant à famille $\mathcal{F}$ donnée. Le résultat est donné dans le théorème suivant.

Théorème 9. *Pour toute fonction ψ appartenant à une famille $\mathcal{F}$, l'inégalité suivante est satisfaite avec une probabilité au moins égale à $1-\epsilon$*

$$R(\psi)\leq R_n(\psi)+C(h,n,\epsilon), \tag{2.5}$$

où C désigne la largeur de l'intervalle de confiance qui dépend de la taille de l'ensemble d'apprentissage n, de la VC-dimension h de $\mathcal{F}$, et du niveau de confiance accordé ϵ, vérifiant l'expression

$$C(h,n,\epsilon)=\sqrt{\frac{h}{n}\left(1-\log 2\frac{h}{n}\right)-\frac{1}{n}\log\frac{\epsilon}{4}}. \tag{2.6}$$

Pour les faibles valeurs du quotient h/n, la largeur C de l'intervalle de confiance est proche de zero, signifiant que le risque empirique $R_n(\psi)$ tend vers le risque réel $R(\psi)$. Il s'agit là de l'essence même de ce principe d'induction. En revanche, pour les valeurs de h/n élevées (proche de 1), la largeur de l'intervalle de confiance est considérable, et $R_n(\psi)$ ne permet plus une bonne estimation de $R(\psi)$. Le compromis entre ces deux quantités est étudié par le principe de minimisation du risque structurel.

2.2.4 Principe de minimisation du risque structurel

Le principe de minimisation du risque empirique incite à la minimisation de celui-ci à tout prix. Toutefois, rien ne garantit que les performances atteintes soient proches du risque optimal, au sens de Bayes. Un contrôle de la capacité en généralisation est souvent nécessaire. Ceci est en particulier indispensable pour les ensembles d'apprentissage de faible taille ayant un quotient h/n supérieur à 5%, où ce quotient correspond à la VC-dimension rapportée par donnée d'apprentissage. Dans ce paragraphe, on présente le principe inductif pour contrôler la capacité en généralisation, en contrôlant la VC-dimension du modèle. C'est le principe de minimisation du risque structurel.

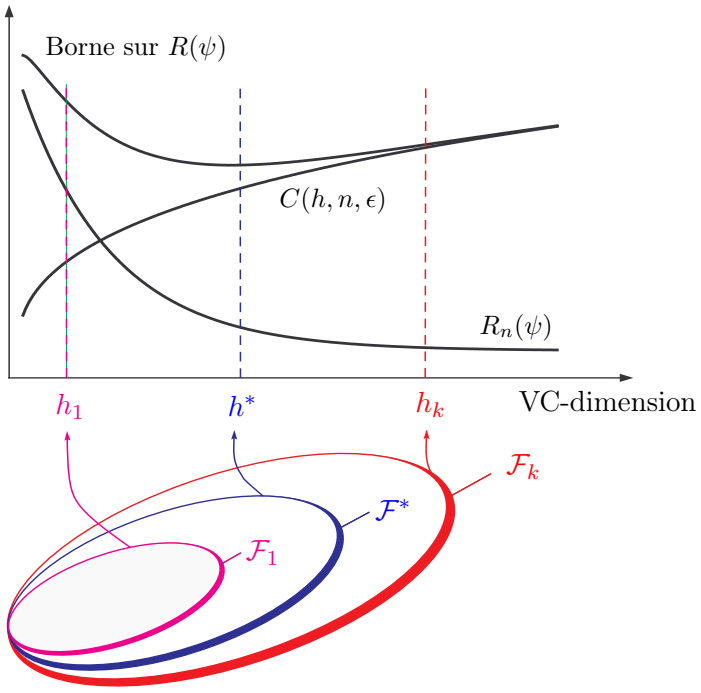


FIGURE 2.3 – Représentation schématique du principe de minimisation du risque structurel.

Le Théorème 9 détermine une borne supérieure sur le risque réel pour toute fonction d'une famille donnée $\mathcal{F}$, avec une probabilité au moins égale à $1 - \epsilon$. La borne supérieure fait intervenir deux termes antagonistes, le risque empirique R_n et la largeur de l'intervalle de confiance C. étant donné un ensemble d'apprentissage, la VC-dimension du modèle permet de contrôler ce compromis comme préconisé par Vapnik avec le principe de minimisation du risque structurel. Pour cela, au lieu d'envisager une seule famille de fonctions, on considère une séquence de plusieurs familles $\mathcal{F}_k$ imbriquées, selon

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_k \subset \ldots,$$

en tenant compte de certaines contraintes techniques énoncées dans [Vap95]. Un exemple particulier d'une telle structure correspond aux familles $\mathcal{F}_k$ de

fonctions polynomiales de degré k. En notant h_k la VC-dimension supposée finie associée à $\mathcal{F}_k$, on peut alors déduire que

$$h_1 \leq h_2 \leq \ldots \leq h_k \leq \ldots.$$

Cette séquence croissante des VC-dimensions a deux conséquences indissociables. On a d'une part une séquence décroissante de risques empiriques correspondant à la fonction optimale de chaque famille, soit

$$\inf_{\psi \in \mathcal{F}_1} R_n(\psi) \geq \inf_{\psi \in \mathcal{F}_2} R_n(\psi) \geq \ldots \geq \inf_{\psi \in \mathcal{F}_k} R_n(\psi) \geq \ldots.$$

D'autre part, on a une séquence croissante du terme relatif à l'intervalle de confiance dans (2.5) tel que

$$C(h_1, n, \epsilon) \leq C(h_2, n, \epsilon) \leq \ldots C(h_k, n, \epsilon) \leq \ldots.$$

En représentant ces résultats sur la Figure 2.3, on résume le principe de minimisation du risque structurel par les deux étapes suivantes :

1. Pour chaque famille $\mathcal{F}_k$, déterminer la fonction optimale minimisant l'erreur empirique
$$\psi^*_{n,\mathcal{F}_k} = \arg\min_{\psi \in \mathcal{F}_k} R_n(\psi).$$
2. Parmi toutes les fonctions optimales obtenues, opter pour celle garantissant la borne supérieure $R_n(\psi^*_{n,\mathcal{F}_k}) + C(h_k, n, \epsilon)$ la plus favorable,
$$\psi^*_n = \arg\min_{k \geq 1} \{R_n(\psi^*_{n,\mathcal{F}_k}) + C(h_k, n, \epsilon)\}.$$

La viabilité théorique de ce principe se heurte à quelques difficultés lors de sa mise en œuvre. Parmi celles-ci, notons la nécessité de connaître la VC-dimension des familles $\mathcal{F}$. De plus, en élaborant un résultat indépendamment de la distribution de probabilité conjointe $P(x, y)$, le Théorème 9 propose une borne supérieure souvent surestimée. Afin de pallier ces inconvénients, des techniques de validation croisée ou de ré-échantillonnage sont souvent considérées, comme étudié par exemple dans [SBSS99] pour les SVM.

2.3 Support vector machines

Au paragraphe précédent, on a posé les fondations de la théorie de l'apprentissage statistique à partir du concept de VC-dimension. Ce paragraphe est dédié à la mise en œuvre de ce cadre théorique à l'aide des SVM. Le critère considéré correspond à la maximisation de la marge entre l'hyperplan séparateur recherché et les éléments de chaque classe de l'ensemble d'apprentissage, comme illustré à la Figure 2.4.

2.3.1 VC-dimension et discrimination linéaire

Depuis leur introduction au début des années 90 dans [BGV92], les SVM ont permis des avancées considérables en reconnaissance des formes, d'une part avec des propriétés de généralisation soutenues par la théorie de l'apprentissage statistique, et d'autre part avec des algorithmes sans cesse plus performants grâce à l'usage du coup du noyau et l'évolution des techniques d'optimisation.

Soit $\mathcal{A}_n$ un ensemble d'apprentissage de n données $x_i \in \mathcal{X}$, avec leurs étiquettes $y_i = \pm 1$. En supposant que ces données sont linéairement séparables, il existe alors une infinité d'hyperplans séparateurs définis par des expressions de la forme $\psi(x) = \langle \boldsymbol{w}, x \rangle_{\mathcal{X}} + w_0$ telles que $y_i \psi(x_i) \geq 1$. On désigne par $\mathcal{F}_\delta$ l'ensemble des hyperplans séparateurs distants d'au moins δ des éléments des deux classes, soit

$$\mathcal{F}_\delta = \{\psi(\cdot) = \langle \boldsymbol{w}, \cdot \rangle_{\mathcal{X}} + w_0 \ : \ |\psi(x)| \geq 1, \|\boldsymbol{w}\| \leq 1/\delta\}.$$

Le théorème suivant est dû à Vapnik [Vap82].

Théorème 10. *La VC-dimension $h_{\mathcal{F}_\delta}$ de la famille $\mathcal{F}_\delta$ est bornée selon la relation*

$$h_{\mathcal{F}_\delta} \leq \min \left\{ \frac{R^2}{\delta^2}, l \right\} + 1,$$

où les données d'apprentissage sont contenues dans une boule de rayon R centrée sur l'origine.

Ce théorème permet la construction d'une séquence de familles imbriquées selon $\mathcal{F}_{\delta_1} \subset \ldots \subset \mathcal{F}_{\delta_k}$, obtenue en considérant différentes valeurs de largeurs de marge regroupées par ordre décroissant. On peut alors obtenir de bonnes propriétés de généralisation en considérant la marge optimale, à partir du principe de la minimisation du risque structurel évoqué à la section précédente.

2.3.2 Cas de données linéairement séparables

Soit $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un ensemble d'apprentissage de données linéairement séparables, représentées par $x_i \in \mathcal{X}$, avec leurs étiquettes $y_i = \pm 1$. Un hyperplan est défini dans cet espace par $\boldsymbol{w}$ et w_0 selon l'expression $\langle \boldsymbol{w}, x \rangle_{\mathcal{X}} + w_0 = 0$. La distance d'un élément x_i à cet hyperplan est alors donnée par

$$\delta_i = \frac{|\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0|}{\|\boldsymbol{w}\|} \tag{2.7}$$

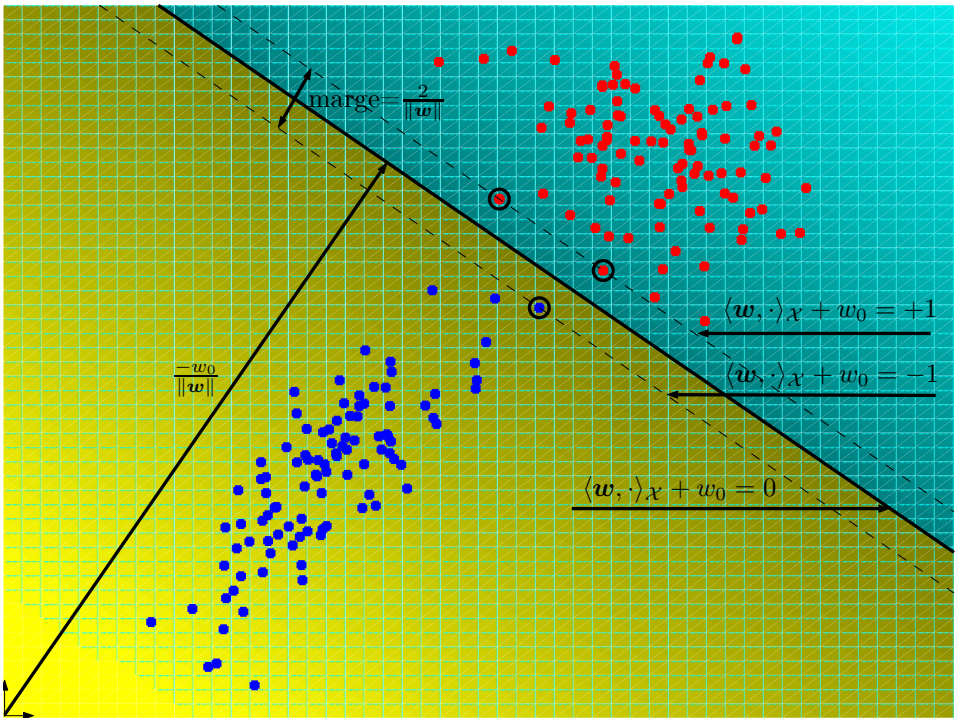


FIGURE 2.4 – Schéma d'un séparateur avec les différentes quantités associées aux SVM. Les support vectors sont identifiés par des cercles

Afin de construire un hyperplan séparateur, on contraint la solution recherchée en imposant

$$\min_{i=1,\dots,n} |\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0| = 1.$$

Celle-ci se traduit par les inégalités suivantes

$$\begin{aligned} \langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0 &\geq +1, \quad \text{si} \quad y_i = +1 \\ \langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0 &\leq -1, \quad \text{si} \quad y_i = -1. \end{aligned}$$

En arrangeant ces relations, on obtient pour tout $i = 1, \dots, n$ la contrainte

$$y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) \geq 1. \tag{2.8}$$

En combinant cette dernière avec l'expression (2.9), on obtient une contrainte sur la distance des éléments de l'ensemble d'apprentissage de l'hyperplan séparateur, selon

$$\delta_i \geq \frac{1}{\|\boldsymbol{w}\|}, \tag{2.9}$$

pour tout $i = 1, \ldots, n$. Par conséquent, l'hyperplan optimal au sens de la marge maximale est donné par la minimisation de $\frac{1}{2}\|\boldsymbol{w}\|^2$ sous la contrainte (2.8). Par l'usage de ce critère convexe, soumis à des contraintes linéaires, la solution optimale est obtenue par le point selle du Lagrangien donné par

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\langle \boldsymbol{w}, x_i\rangle_{\mathcal{X}} + w_0) - 1\big), \tag{2.10}$$

où $\alpha_1, \ldots, \alpha_n \geq 0$ désignent les multiplicateurs de Lagrange. Ce point-selle correspond à l'annulation des dérivées partielles par rapport aux différentes variables recherchées et aux multiplicateurs de Lagrange. L'optimalité de ceux-ci étant désignée par $\boldsymbol{w}^*$, w_0^* et α_i^*, on aboutit alors aux conditions suivantes :

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \tag{2.11}$$

$$\sum_{i=1}^{n} \alpha_i^* y_i = 0 \tag{2.12}$$

En injectant ces deux résultats dans l'expression du Lagrangien (2.10), le problème d'optimisation se traduit alors par la maximisation de la forme duale du Lagrangien

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \ \langle x_i, x_j\rangle_{\mathcal{X}}, \tag{2.13}$$

sous la contrainte $\sum_{i=1}^{n} \alpha_i y_i = 0$, ainsi que la positivité des multiplicateurs de Lagrange α_i. On sait que les coefficients ainsi obtenus vérifient la condition de Kühn-Tucker

$$\alpha_i \big(y_i(\langle \boldsymbol{w}, x_i\rangle_{\mathcal{X}} + w_0) - 1\big) = 0, \tag{2.14}$$

pour $i = 1, \ldots, n$. En conséquence, seules les données x_i satisfaisant l'égalité dans la contrainte (2.8), $y_i(\langle \boldsymbol{w}, x_i\rangle_{\mathcal{X}} + w_0) = 1$, admettent des multiplicateurs de Lagrange α_i non-nuls. Ces données vérifient $\delta_i = 1$, ce qui signifie qu'il n'existe pas de donnée plus proche de l'hyperplan optimal que x_i. De tels échantillons sont appelés *support vectors* puisqu'ils définissent à eux seuls l'hyperplan optimal, caractérisé par $\boldsymbol{w}^*$ et w_0^*. Le premier est donné par l'expression (2.11), les échantillons correspondant à α_i non-nul n'y contribuant pas. Le second paramètre w_0^* permet de situer l'hyperplan à mi-distance des deux classes. On exprime alors le seuil selon

$$w_0^* = \frac{1}{2}\big(\langle \boldsymbol{w}^*, x_j\rangle_{\mathcal{X}} + \langle \boldsymbol{w}^*, x_k\rangle_{\mathcal{X}}\big),$$

où x_j et x_k désignent deux support vectors appartenant à deux classes différentes. Plus généralement, les support vectors regroupent l'information discriminante de l'ensemble d'apprentissage. Les autres échantillons pourraient être retirés de ce dernier et la répétition de l'algorithme à partir des support vectors uniquement produirait le même hyperplan optimal.

Quelques avantages de l'approche SVM

Les support vectors contribuent à l'attrait pour les SVM en raison du caractère parcimonieux des solutions obtenues. Ceci est dû au critère de maximisation de la marge, qui a par ailleurs d'autres effets positifs pour l'approche SVM. On a d'une part un problème d'optimisation quadratique, n'admettant donc pas de minima local mais une solution unique si la matrice formée par les quantités $y_i y_j \langle x_i, x_j \rangle_{\mathcal{X}}$ est de rang plein. D'autre part, le résultat obtenu est relativement robuste aux faibles variations des paramètres en question. Ces avantages constituent autant d'avancées considérables pour les méthodes de reconnaissance des formes classiques. L'approche des réseaux de neurones par exemple nécessite la résolution d'un problème d'optimisation admettant des optimum locaux. De plus, la parcimonie de la solution nécessite une étape supplémentaire d'élagage. Au-delà de ces propriétés intéressantes, on rappelle que la maximisation de la marge est motivée par le principe de minimisation du risque structurel. Celui-ci correspond à la recherche d'un compromis entre l'erreur empirique et la richesse de l'ensemble des fonctions candidates, comme illustré à la Figure 2.3. Les deux classes étant linéairement séparables, on se restreint à l'ensemble des hyperplans séparateurs, produisant donc une erreur empirique nulle. Parmi ceux-ci, on cherche celui avec les meilleurs propriétés de généralisation, mesurée par une faible VC-dimension conformément au Théorème 10.

Plus généralement, le principe de minimisation du risque structurel s'applique dans le cas de classes non-linéairement séparables. La mise en œuvre des SVM pour ce type de problèmes est traité dans la section suivante.

2.3.3 Cas de classes non-linéairement séparables

Précédemment, on s'est restreint aux problèmes où les données d'apprentissage sont linéairement séparables. Ceci n'est pas le cas en général. On peut toutefois considérer un hyperplan séparant les deux classes en question, en tolérant des erreurs de classification sur certains échantillons. L'erreur de classification de l'échantillon x_i est quantifiée par les variables positives de pénalisation ξ_i, dites *slack variables* ou variables de relaxation, ce qui permet

de généraliser les contraintes dans (2.8) ainsi

$$y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) \geq 1 - \xi_i, \tag{2.15}$$

pour tout $i = 1, \ldots, n$, avec $\xi_1, \ldots, \xi_n \geq 0$. La minimisation de l'erreur totale $\sum_{i=1}^{n} \xi_i$, combinée avec le critère de la la marge, correspond à la minimisation du critère

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i,$$

sous les contraintes (2.15) avec C un paramètre permettant une pondération relative des deux termes antagonistes : une solution très régulière est obtenue pour $C = 0$ tandis qu'on aboutit à une classification parfaite pour $C = \infty$. En d'autres termes, il s'agit d'un paramètre de régularisation permettant un contrôle supplémentaire de la capacité du classifieur résultant. On retrouve la même formulation que celle de la théorie de la régularisation selon Tikhonov présentée dans la Section 1.1.2, avec $\eta = 1/C$.

La solution de ce problème d'optimisation avec contraintes est donnée par le point selle du Lagrangien

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \big(y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) - 1 + \xi_i\big) - \sum_{i=1}^{n} \beta_i \xi_i,$$

α_i et β_i désignant les multiplicateurs de Lagrange. La minimisation du Lagrangien par rapport à $\boldsymbol{w}$, w_0 et les ξ_i produit les conditions suivantes :

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

$$\sum_{i=1}^{n} \alpha_i^* y_i = 0$$

$$\alpha_i^* + \beta_i^* = C$$

Le problème dual est alors obtenu en injectant ces expressions dans la forme primaire, et la solution est donnée par

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \ \langle x_i, x_j \rangle_{\mathcal{X}} - \sum_{i=1}^{n} \alpha_i, \tag{2.16}$$

sous les contraintes

$$0 \geq \alpha_i \geq C \qquad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

La résolution de ce problème est similaire à celle obtenue dans le cas de classes linéairement séparables, avec une contrainte supplémentaire sur les α_i.

2.4 SVM dans un RKHS

Les différentes expressions montrent que la méthode présentée dans les Sections 2.3.2 et 2.3.3 se prête aisément à une généralisation dans un espace transformé à noyau reproduisant. On considère le noyau reproduisant κ et l'espace de Hilbert associé $\mathcal{H}$, l'échantillon x_i est alors représenté dans cet espace par κ_{x_i}, et le produit scalaire est désigné par $\kappa(x_i, x_j)$, pour tout $x_i, x_j \in \mathcal{X}$. La forme duale du Lagrangien dans (2.13) ou (2.16), s'écrit alors

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \ \kappa(x_i, x_j), \tag{2.17}$$

sous des contraintes inchangées. La résolution de ce problème d'optimisation avec contrainte détermine les multiplicateurs de Lagrange, et par conséquent l'hyperplan optimal dans l'espace fonctionnel $\mathcal{H}$, défini par les paramètres

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \kappa_{x_i} \tag{2.18}$$

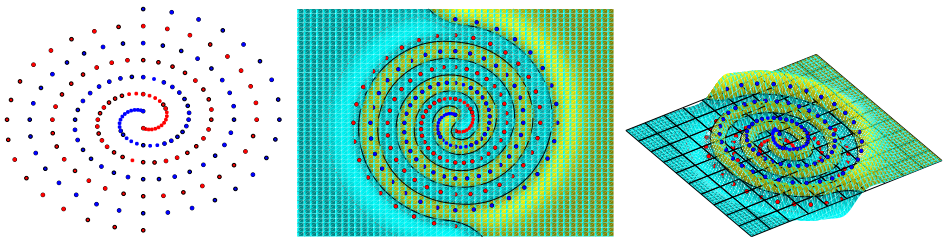
et le seuil par

$$w_0^* = \frac{1}{2} \sum_{i=1}^{n} \alpha_i^* y_i \big(\kappa(x_i, x_j) + \kappa(x_i, x_k) \big), \tag{2.19}$$

où x_j et x_k désignent deux support vectors appartenant à deux classes différentes. On retrouve les effets du Théorème de Représentation dans l'expression de $\boldsymbol{w}^*$, et ceux de la propriété reproduisante dans l'expression du seuil. La règle de décision consiste à comparer la statistique $\langle \boldsymbol{w}^*, \kappa_x \rangle_{\mathcal{H}}$ au seuil w_0^*, soit

$$\Lambda(x) = \sum_{i=1}^{n} \alpha_i^* y_i \kappa(x, x_i) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} \sum_{i=1}^{n} \alpha_i^* y_i \big(\kappa(x_i, x_j) + \kappa(x_i, x_k) \big).$$

L'évaluation de la règle de décision ne nécessite à aucun moment d'exhiber l'espace de représentation $\mathcal{H}$, ni les fonctions noyau de l'ensemble de données



(a) (b) (c)

FIGURE 2.5 – Résultat obtenu par les SVM pour la résolution du problème de classification de deux classes sous forme de spirales enroulées l'une autour de l'autre.

disponibles, une fois que leur produit scalaire est connu. Afin d'illustrer l'efficacité des SVM avec un noyau reproduisant non-linéaire, on considère le problème classique de classification de deux classes prenant la forme de spirales enroulées l'une autour de l'autre, voir Figure 2.5 (a). Un classifieur linéaire ne peut évidemment pas mener à une discrimination satisfaisante. On considère alors le noyau Gaussien, défini par $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma_0^2)$ où σ_0 est le paramètre de largeur de bande. Comme illustré à la Figure 2.5 $(b)-(c)$, ce noyau permet une discrimination entre les deux classes.

Annexe A

Noyaux (reproduisants) classiques

	Noyau [1]	Expression mathématique
Noyau projectif	monomial	$(\langle x_i, x_j \rangle)^p$
	polynomial	$(c + \langle x_i, x_j \rangle)^p$
	polynomial de Vovk	$\frac{1-\langle x_i, x_j \rangle^p}{1-\langle x_i, x_j \rangle}$
	polynomial inifini	$(1 - \langle x_i, x_j \rangle)^{-p}$
	exponentiel	$\exp(\langle x_i, x_j \rangle / 2\sigma^2)$
	sigmoïde (perceptron)	$\tanh(\langle x_i, x_j \rangle / \sigma + c)$
Noyau radial	Gaussien	$\exp(-\|x_i - x_j\|^2 / 2\sigma^2)$
	Laplacien	$\exp(-\|x_i - x_j\| / 2\sigma^2)$
	multiquadrique	$(\|x_i - x_j\|^2 + c^2)^{1/2}$
	multiquadrique inverse	$(\|x_i - x_j\|^2 + c^2)^{-1/2}$
	splines (*thin plate*)	$\|x_i - x_j\|^{2n+1}$
	splines log (*thin plate*)	$\|x_i - x_j\|^{2n} \log(\|x_i - x_j\|)$
	B-splines sur $\mathbb{R}$	$B_{2n+1}(x_i - x_j)$
	trigonométrique polynomial sur $\mathbb{R}$	$\frac{\sin(d+1/2)(x_i - x_j)}{\sin(x_i - x_j)/2}$
	de Fourier sur $\mathbb{R}$ (régularisation faible)	$\frac{\pi}{2\gamma} \frac{\cosh \frac{\pi - \|x_i - x_j\|}{\gamma}}{\sinh \pi/\gamma}$
	de Fourier sur $\mathbb{R}$ (régularisation forte)	$\frac{1-\gamma^2}{2(1-2\gamma \cos(x_i - x_j) + \gamma^2)}$
	Anova 1	$\left(\sum_{k=1}^n \exp((x_i(k) - x_j(k))^2 \theta)\right)^2$
	quadratique rationnel	$1 - \frac{\|x_i - x_j\|^2}{c + \|x_i - x_j\|^2}$
	circulaire sur $\mathbb{R}^2$	$\frac{2}{\pi} \arccos \frac{\|x_i - x_j\|}{\theta} - \frac{2}{\pi} \frac{\|x_i - x_j\|}{\theta} \sqrt{1 - \frac{\|x_i - x_j\|^2}{\theta^2}}$
	sphérique sur $\mathbb{R}^3$	$1 - \frac{3}{2}\frac{\|x_i - x_j\|}{\theta} + \frac{1}{2}\left(\frac{\|x_i - x_j\|}{\theta}\right)^3$
	onde $\mathbb{R}^3$	$\frac{\theta}{\|x_i - x_j\|} \sin \|x_i - x_j\| / \theta$

1. Certains noyaux ne sont pas définis positifs que pour des valeurs particulières de leurs paramètres, comme c'est le cas de la fonction sigmoïde.

Annexe B

Méthodes à noyaux les plus connues

	Algorithme	Fonction coût $V(\psi(x_i), y_i)$
Non-supervisés	Estimation de densité [Vap95]	$-\log(\psi(x_i))$
	Analyse en composantes principales [VTS04]	$-\lvert\psi(x_i)\rvert^2/\lVert\psi\rVert^2_{\mathcal{H}}$
	Projection pursuit [Sun98]	$\text{skewness}\{\psi(x_i)\}, \text{kurtosis}\{\psi(x_i)\}$, $\text{entropie}\{\psi(x_i)\}, \ldots$
	Variétés principales régularisées [SWMS99]	
	Detection de nouveauté (ν-SVM) [KSW04, SWS$^+$00]	$\max\{\rho - \psi(x_i), 0\} - \nu\rho$
Moindres carrés	Régression ridge [SGV98, HTF01]	$(y_i - \psi(x_i))^2$
	Least square support vector machine [SV99, SGB$^+$02]	$(y_i - \psi(x_i))^2$
	Regularized least square classification [CB04, Rif02]	$(y_i - \psi(x_i))^2$
	Réseaux de régularisation [EPP99]	$(y_i - \psi(x_i))^2$
	Proximal support vector machine [FM01]	$(y_i - \psi(x_i))^2$
	Moindres carrés modifiés [Zha04]	$\max\{y_i - \psi(x_i), 0\}^2$
Régression	Support Vector Regression [Vap95]	
	ϵ-insensitive	$\max\{\lvert y_i - \psi(x_i)\rvert - \epsilon, 0\}$
	ϵ-insensitive quadratique	$\max\{\lvert y_i - \psi(x_i)\rvert - \epsilon, 0\}^2$
	Régression en quantile [TLSS06]	$\tau(y_i - \psi(x_i))$ si $y_i \geq \psi(x_i)$ $(\tau - 1)(y_i - \psi(x_i))$ sinon
	AdaBoost [Vap95, HTF01]	$\exp(-y_i\psi(x_i))$
	Régression robuste de Huber [Hub81, Vap95]	$\frac{1}{4\epsilon}\lvert y_i - \psi(x_i)\rvert^2$ si $\lvert y_i - \psi(x_i)\rvert \leq 2\epsilon$ $\lvert y_i - \psi(x_i)\rvert - \epsilon$ sinon
Classification	Support Vector Classification : [Vap95]	
	Marge dure (indicateur)	$\mathbb{1}_{1-y_i\psi(x_i)}$
	Mal-classification (indicateur)	$\mathbb{1}_{-y_i\psi(x_i)}$
	Marge diffuse (hinge)	$\max\{1 - y_i\psi(x_i), 0\}$
	Marge diffuse (mal-classification)	$\max\{-y_i\psi(x_i), 0\}$
	Marge diffuse (hinge) quadratique	$\max\{1 - y_i\psi(x_i), 0\}^2$
	Régression logistique (à noyaux)	$\log(1 + \exp(-y_i\psi(x_i)))$
	Import Vector Machine [ZH02]	$y_i\psi(x_i) - \log(1 + \exp(\psi(x_i)))$
	ν-SVM [KSW04]	$\max\{\rho - y_i\psi(x_i), 0\} - \nu\rho$

Bibliographie

[Abe03] S. Abe. Analysis of multiclass support vector machines. In *Proc. International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'2003)*, pages 385–396, Vienna, Austria, 2003.

[ABR64] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25 :821–837, 1964.

[Aro50] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68 :337–404, 1950.

[BGV92] B. Boser, I. Guyon, and V.N. Vapnik. An training algorithm for optimal margin classifiers. In *Proc. 5th Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[Cap06] A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical Report Technical Report MIT-CSAIL-TR-2006-062, Computer Science and Artificial Intelligence Laboratory, massachusetts institute of technology, Cambridge, MA, USA, September 2006.

[CB04] N. Cesa-Bianchi. Applications of regularized least squares to classification problems. In S. Ben-David, J. Case, and A. Maruoka, editors, *Proc. 15th International Conference ALT 2004 : Algorithmic Learning Theory, October 2-5, 2004*, volume 3244, pages 14–18, Padova, Italy, 2004. Springer.

[CH02] Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Comput.*, 14(12) :2791–2846, 2002.

[Cov65] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14 :326–334, 1965.

[CS00] K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In *Proc. the Thirteenth*

Annual Conference on Computational Learning Theory (COLT '00), pages 35–46, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[CS02a] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2 :265–292, 2002.

[CS02b] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory : on the bias-variance problem. *Found. Comput. Math.*, 2(4) :413–428, 2002.

[CS02c] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1) :1–49, 2002.

[EPM00] A. Elisseeff and H. Paugam-Moisy. A new multi-class svm based on a uniform convergence result. In *Proc. the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)*, volume 4, page 4183, Washington, DC, USA, 2000. IEEE Computer Society.

[EPP99] T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1999.

[FM01] G. Fung and O.L. Mangasarian. Proximal support vector machine classifiers. In *Proc. seventh ACM SIGKDD International Conference on Knowledge Discovery and data mining (KDD '01)*, pages 77–86, New York, NY, USA, 2001. Asscociation for Computing Machinery Press.

[Her02] R. Herbrich. *Learning kernel classifiers. Theory and algorithms.* The MIT Press, Cambridge, MA, USA, 2002.

[HTF01] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning.* Springer, New York, August 2001.

[Hub81] P.J. Huber. *Robust Statistics.* Wiley, New York, NY, USA, 1981.

[Iva76] V.V. Ivanov. The theory of approximate methods and their application to the numerical solution of singular integral equations (translated from the Russian by A. Ideh). In R.S. Anderssen and D. Elliott, editors, *Monographs and Textbooks on Mechanics of Solids and Fluids, Mechanics : Analysis*, 2, Leyden, Netherlands, 1976. Noordhoff International Publishing.

[Kai71] T. Kailath. RKHS approach to detection and estimation problems–i : Deterministic signals in gaussian noise. *IEEE Transactions on Information Theory*, 17(5) :530–549, September 1971.

[Kre99] U.H.-G. Kreßel. Pairwise classification and support vector machines. *Advances in kernel methods : support vector learning*, pages 255–268, 1999.

[KSW04] J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Trans. Signal Processing*, 52(8), Aug 2004.

[Kur04] V. Kurková. Learning from data as an inverse problem. In J. Antoch, editor, *Proc. Computational Statistics (CompStat'04)*, pages 1377–1384, Heidelberg,Germany, 2004. Physica-Verlag/SpringerAcademic Press, Inc.

[KW71] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33 :82–95, 1971.

[Mac03] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. Available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

[MNY06] H.Q. Minh, P. Niyogi, and Y. Yao. Mercer's theorem, feature maps, and smoothing. In *Proc. 19th Annual Conference on Learning Theory (COLT 2006)*. Springer, 2006.

[MR97] P. Maaß and A. Rieder. Wavelet-accelerated tikhonov–phillips regularization with applications. In A.K. Louis and W. Rundell, editors, *Inverse Problems in Medical Imaging and Nondestructive Testing*, pages 134–158, Springer, Vienna, 1997.

[Muk04] S. Mukherjee. Statistical learning : Algorithms and theory. Course notes for STA270 : statistical methods for computational biology, Institute of Statistics and Decision Sciences (ISDS), Duke University, Durham, NC, USA, November 2004.

[Par70] E. Parzen. Statistical inference on time series by RKHS methods. In R. Pyke, editor, *Proc. 12th Biennial Seminar*, pages 1–37, Montreal, Canada, 1970. Canadian Mathematical Congress.

[PG90] T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78 :1481–1497, 1990.

[Phi62] D.L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *J. ACM*, 9(1) :84–97, 1962.

[Rif02] R.M. Rifkin. *Everything Old Is New Again : A Fresh Look at Historical Approaches in Machine Learning*. Phd thesis, Sloan School of Management Science : Massachusetts Institute of Technology, September 2002.

[SABO99] S. Saitoh, D. Alpay, J.A. Ball, and T. Ohsawa, editors. *Reproducing Kernels and Their Applications*, volume 3, Dordrecht, The Netherlands, 1999. International Society for Analysis, Applications and Computation (ISAAC'97), Kluwer Academic Publishers.

[SBSS99] A.J. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, USA, 1999.

[SGB+02] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002.

[SGV98] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proc. of the Fifteenth International Conference on Machine LearningICML '98*, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[SHW00] B. Schölkopf, R. Herbrich, and R. Williamson. A generalized representer theorem. Technical Report NC2-TR-2000-81, NeuroCOLT, Royal Holloway College, University of London, UK, 2000.

[SS01] B. Schölkopf and A.J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[STC04] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[Ste02] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2 :67–93, 2002.

[Sun98] J. Sun. Projection pursuit. In S. Kotz, C. Read, D. Banks, and N. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 2, pages 554–560. Wiley, 1998.

[SV99] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3) :293–300, 1999.

[SWMS99] A. Smola, R.C. Williamson, S. Mika, and B. Schölkopf. Regularized principal manifolds. In *Computational Learning Theory : 4th European Conference*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 214–229. Springer, 1999.

[SWS+00] B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, and J.C. Platt. Support vector method for novelty detection. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 582–588, Cambridge, MA, USA, 2000. MIT Press.

[TA77] A.N. Tikhonov and V.Y. Arsenin. *Solutions of ill-posed problems*. John Wiley, New York, 1977.

[Tik63] A.N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Sov. Math. Dokl.*, 4 :1035–1038, 1963.

[TLSS06] I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola. Nonparametric quantile regression. *Journal of Machine Learning Research*, 7 :1231–1264, 2006.

[Vap82] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer-Verlag New York, Inc., Secaucus, NJ, USA, springer edition, 1982.

[Vap95] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, September 1998.

[Vay00] N. Vayatis. *Inégalités de Vapnik-Chervonenkis et mesures de complexité*. Ph.d. thesis, Ecole Polytechnique, Palaiseau, France, jan 2000. in english.

[VC71] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2) :264–280, 1971.

[VTS04] J.P. Vert, K. Tsuda, and B. Schölkopf. A primer on kernel methods. In B. Schölkopf, K. Tsuda, and J.P. Vert, editors, *Kernel Methods in Computational Biology*, pages 35–70, Cambridge, MA, USA, 2004. MIT Press.

[Wah90] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Math (SIAM), Philadelphia, 1990.

[WLZ00] G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross validation for support vector machines. In A. Smola, P. Bartlett,

B. Schölkopf, and C. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–309, Cambridge, MA, USA, 2000. MIT Press.

[ZH02] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS)*, Cambridge, MA, USA, 2002. MIT Press.

[Zha04] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32 :56–134, March 2004.