Introduction aux fonctions de décision non-linéaires

Thierry Bazier-Matte

7 avril 2017

1 Introduction aux fonctions de décisions non linéaires

Ce chapitre se veut une brève introduction aux propriétés des espaces de décision obtenus par noyaux reproduisants. En premier lieu, une discussion sur la forme duale du problème linéaire ainsi que les propriétés des espaces à noyau permettront d'obtenir une meilleure intuition (Section 1.1). Par la suite, la Section 1.2 présentera quels algorithmes permettant de trouver une politique d'investissement optimal à partir d'un ensemble d'entraînement $\mathcal{S}_n = \{x_i, r_i\}_{i=1}^n \sim M^n$ échantilloné à partir de la distribution de marché et d'une fonction d'utilité concave. Quelques exemples de noyaux courants seront présentés, suivis des dérivations des deux formes d'optimisation.

1.1 Propriétés des espaces de décision à noyau reproduisant

Formulations primales et duales Tel que discuté en introduction, le cas le plus simple pour un espace de décision $\boldsymbol{Q}$ est celui où $\boldsymbol{Q} = \boldsymbol{X}^*$, c'est-à-dire le dual de l'espace vectoriel $\boldsymbol{X}$ [1]. La *décision* prise suite à l'observation d'un vecteur d'information $x \in \boldsymbol{X}$ est simplement $q(x) = q^T x$. Le problème à résoudre est ainsi

$$\underset{q \in \mathscr{R}^p}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q^T x_i) - \lambda \|q\|^2, \tag{1}$$

duquel on tire un $\hat{q}$ optimal. Cette formulation *primale* est intuitivement claire : on cherche à maximiser l'utilité moyenne suivant une politique unique q appliquée à chaque observation x_i, tout en cherchant à éviter de favoriser excessivement une des dimensions d'information par rapport aux autres. Or, selon le théorème de la représentation qui sera présenté un peu plus loin (p. 6), la politique optimale $\hat{q}$ peut également s'exprimer comme une combinaison linéaire des observations x_i. Ainsi, en notant

1. Le *dual* $\boldsymbol{V}^*$ d'un espace vectoriel $\boldsymbol{V}$ correspond à l'ensemble des formes linéaires sur $\boldsymbol{V}$. Dans le cas fini où $\boldsymbol{V} = \mathscr{R}^m$, alors un élément $w^* \in \boldsymbol{V}^*$ est souvent représenté par un vecteur ligne w^T à m éléments, tel que $w^*(v) = w^T v$.

$\Xi \in \mathscr{R}^{n \times p}$ la matrice des n observations de x, il existe $\hat{\alpha} \in \mathscr{R}^n$ tel que

$$\hat{q} = \Xi^T \hat{\alpha}. \tag{2}$$

Cette propriété fondamentale permet donc de chercher une combinaison linéaire optimale $\hat{\alpha} \in \mathscr{R}^n$ à partir de laquelle la politique optimale peut être déduite. En substituant (2) dans (1), on obtient la *représentation duale* du problème :

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, \alpha^T \Xi x_i) - \lambda \alpha^T \Xi \Xi^T \alpha. \tag{3}$$

Si, à des fins de simplification d'interprétation, l'investisseur est neutre au risque, et en notant $K \coloneqq \Xi \Xi^T \in \mathscr{R}^{n \times n}$, *i.e.*, $K_{ij} = x_i^T x_j$, alors le problème sous sa forme duale s'exprime comme

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \alpha^T K r - \lambda \alpha^T K \alpha. \tag{4}$$

Intuitivement, la matrice K, étant semi-définie positive, représente une *covariance de similarité* entre chacune des observations x_i, où la variance de chaque observation est donnée par sa norme $\|x_i\|^2$ et la corrélation entre deux observations par le cosinus de l'angle : $\rho_{ij} = x_i^T x_j / \|x_i\| \|x_j\|$. L'expression $n^{-1} K r \in \mathscr{R}^n$ indique quelles dimensions permettent d'obtenir le meilleur rendement en considérant l'influence pondérée de toutes les observations :

$$[Kr]_j = n^{-1} \sum_{i=1}^{n} r_i \rho_{ij} \|x_i\| \|x_j\|. \tag{5}$$

Le rôle de α est alors de choisir les dimensions les plus favorables ; enfin le terme de régularisation $\lambda \alpha^T K \alpha$ a pour effet non seulement de choisir une solution finie (puisque quadratique), mais aussi de standardiser l'effet de chaque dimension afin de limiter par exemple l'influence d'observations dotées d'une norme plus élevée que les autres.

On note finalement que la solution analytique du problème risque neutre devient

$$K\alpha = \frac{1}{2n\lambda} Kr. \tag{6}$$

entraînant sans surprise $\hat{\alpha} = (2n\lambda)^{-1} r$ si K est de plein rang. Si par contre K n'est pas de plein rang, c'est-à-dire s'il existe une observation de norme nulle ($\|x_i\| = 0$) ou colinéaire par rapport à une autre ($x_i = k x_j$ entraîne $\rho_{ij} = 1$), $\hat{\alpha}$ n'est pas défini puisqu'il existe alors une infinité de solutions. Il est à noter que le théorème de la représentation n'est pas forcément *nécessaire*, il est simplement suffisant.

Nous verrons cependant une autre forme duale au problème dont la solution $\hat{\alpha}$ est en bijection avec $\hat{q}$.

Décisions non-linéaires Si cette classe des décisions linéaires a l'avantage d'être simple, elle est en revanche fort peu adaptée à des situations pourtant peu complexes.

Géométriquement, elle ne fait que séparer l'espace $\boldsymbol{X}$ en deux : un côté entraînera des décisions d'investissement positifs, l'autre des décisions négatives. **[Todo:** Problème XOR irrésoluble**]**.

La méthode des noyaux permet de circonvenir ce problème en remplaçant la notion de similarité entre deux points par une fonction de noyau semi-défini positif κ.

Définition. Un *noyau semi-défini positif*, ou simplement un *noyau* $\kappa : \boldsymbol{X} \times \boldsymbol{X} \to \mathscr{R}$ est tel que pour tout ensemble $\{x_1, \dots, x_n\} \in \boldsymbol{X}^n$, la matrice $K_{ij} = \kappa(x_i, x_j)$ est semi-définie positive.

Proposition 1. *Tout noyau semi-défini positif κ induit un espace de décision $\boldsymbol{Q}$ [2] doté d'un produit scalaire $\langle \cdot, \cdot \rangle : \boldsymbol{Q} \times \boldsymbol{Q} \to \mathscr{R}$ ainsi que d'une application $\phi : \boldsymbol{X} \to \boldsymbol{Q}$ donnée par $\phi(x) = \kappa(x, \cdot) = \kappa(\cdot, x)$. De plus, $\boldsymbol{Q}$ dispose de la propriété reproductive par laquelle pour tout $q \in \boldsymbol{Q}$, $q(x) = \langle q, \phi(x) \rangle$. En particulier on en conclut que $\kappa(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$. Finalement, l'inégalité de Cauchy-Swartz s'applique à $\boldsymbol{Q}$: pour tout $q_1, q_2 \in \boldsymbol{Q}$, $\langle q_1, q_2 \rangle^2 \leq \|q_1\| \|q_2\|$, où la norme de q est définie par $\|q\|^2 = \langle q, q \rangle$. En particulier, on note que $q(x)^2 \leq \|q\| \kappa(x, x)$.*

Ainsi, doté d'un noyau κ, on obtient un espace de décision $\boldsymbol{Q}$ tel que le problème primal s'exprime par

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \, q(x_i)) - \lambda \|q\|^2. \tag{7}$$

Il convient de noter que chaque type de noyau entraîne une classe de décision bien particulière. Ainsi, selon la géométrie de la densité de la distribution M, certains noyaux seront plus adaptés que d'autre. D'une certaine façon, il s'agit là d'une faiblesse du modèle car celui-ci est incapable de *déterminer* le bon noyau à employer et cette tâche revient alors au gestionnaire de portefeuille.

Exemples Outre le *noyau linéaire*, défini par $\kappa(x_1, x_2) = x_1^T x_2$, les *noyaux polynômiaux d'ordre k* donnés par $\kappa(x_1, x_2) = (x_1^T x_2 + c)^k$ sont également courants. Ces types de noyaux ont cependant l'inconvénient de conserver une notion d'amplitude absolue ; on peut à l'inverse définir des noyaux invariants au déplacement et à la rotation, *i.e.* tels que $\kappa(x_1, x_2) = \kappa(\|x_1 - x_2\|)$. La notion de similarité ne dépend alors plus que de la distance entre deux points. Ainsi, le noyau gaussien κ_σ sera défini par :

$$\kappa_\sigma(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2} \right), \tag{8}$$

où σ représente la sensibilité du noyau ; des valeurs élevées de σ le rendront rapidement insensible à des données pourtant rapprochées dans l'espace $\boldsymbol{X}$ alors qu'une valeur σ faible leur accordera une similarité beaucoup plus grande.

2. Pour être tout à fait exact, $\boldsymbol{Q}$ est alors un espace de Hilbert à noyau reproduisant.

Enfin, ces noyaux peuvent se recombiner afin d'en former de nouveaux. Voir Bishop et Mohri.

1.2 Algorithmes de décision non-linéaires

Magré que le problème primal soit bien posé, l'espace $\boldsymbol{Q}$ est a priori inconnu et peut de surcroit être de dimension infinie. Il est donc nécessaire de déterminer une méthode algorithmique capable de déterminer $\hat{q}$. Si la matrice de similarité K est définie positive, alors on peut utiliser le théorème de la représentation pour résoudre le problème suivant :

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad \sum_{i=1}^{n} u(r_i\, \alpha^T \psi(x_i)) - \alpha^T K \alpha. \tag{9}$$

où $\psi : \boldsymbol{X} \to \mathscr{R}^n$ est un opérateur linéaire tel que $\psi(x_i)_j = \kappa(x_i, x_j)$; c'est en fait la contrepartie de l'application de Ξ sur x_i dans le cas linéaire. Par ailleurs la décision optimale s'exprime comme $\hat{q} = \hat{\alpha}^T \psi$, c'est-à-dire comme une combinaison linéaire de fonctions non-linéaires. Enfin, si K n'est pas définie positive, il suffit alors de ne considérer que les points sans co-linéarité ou avec norme non nulle.

Il existe aussi une autre façon de résoudre le problème primal qui consiste à dualiser le problème primal dans le cas linéaire pour voir émerger la matrice de similarité K, ce qui permet alors de considérer n'importe quel noyau. Par ailleurs, cette méthode est valide que K soit de plein rang ou non. Ainsi, le problème primal peut se résoudre suivant le problème

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad -\sum_{i=1}^{n} \ell^*(\alpha_i / r_i) - \frac{1}{4n\lambda} \alpha^T K \alpha. \tag{10}$$

La fonction $\ell^* : \mathscr{R} \to \mathscr{R}$ est le *conjugué convexe* de la fonction de perte $\ell = -u$ (voir (21), p. 5). La décision est donnée par

$$q(x) = -\frac{1}{2n\lambda} \alpha^T \psi(x). \tag{11}$$

Par exemple, dans le cas d'une utilité risque neutre, $\ell^* = \infty$ sauf si $\alpha_i / r_i = -1$, donc nécessairement $\alpha = -r$ et alors

$$q(x) = \frac{1}{2n\lambda} r^T \psi(x). \tag{12}$$

1.3 Démonstrations

Approche duale On cherche à résoudre le problème suivant, avec $q \in \mathscr{R}^p$ comme variable d'optimisation :

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell(r_i\, q^T x_i) + n\lambda \|q\|^2, \tag{13}$$

où $\ell = -u$. De façon équivalente, en introduisant un nouveau vecteur $\xi \in \mathscr{R}^n$, on a

$$\begin{aligned} \text{minimiser} \quad & \sum_{i=1}^{n} \ell(\xi_i) + n\lambda\|q\|^2 \\ \text{tel que} \quad & \xi_i = r_i\, q^T x_i. \end{aligned} \tag{14}$$

Soit $\alpha \in \mathscr{R}^n$. Le lagrangien de (14) peut s'exprimer comme

$$\mathscr{L}(q, \xi, \alpha) = \sum_{i=1}^{n} \ell(\xi_i) + n\lambda\|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i). \tag{15}$$

Puique l'objectif de (14) est convexe et que ses contraintes sont affines en q et ξ, on peut appliquer le théorème de Slater qui spécifie que le saut de dualité du problème est nul. En d'autres mots, résoudre (13) revient à maximiser la fonction dual de Lagrange g sur α :

$$\text{maximiser} \quad g(\alpha) = \inf_{q,\xi} \mathscr{L}(q, \xi, \alpha). \tag{16}$$

On note que

$$g(\alpha) = \inf_{q,\xi} \left\{ \sum_{i=1}^{n} \ell(\xi_i) + n\lambda\|q\|^2 + \sum_{i=1}^{n} \alpha_i (r_i q^T x_i - \xi_i) \right\} \tag{17}$$

$$= \inf_{\xi} \left\{ \sum_{i=1}^{n} \ell(\xi_i) - \alpha^T \xi \right\} + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda\|q\|^2 \right\} \tag{18}$$

$$= -\sup_{\xi} \left\{ a^T \xi - \sum_{i=1}^{n} \ell(\xi_i) \right\} + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda\|q\|^2 \right\} \tag{19}$$

$$= -\sum_{i=1}^{n} \ell^*(\alpha_i) + \inf_{q} \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda\|q\|^2 \right\}. \tag{20}$$

Où ℓ^* est le conjugé convexe de la fonction de perte et est définie par

$$\ell(\alpha_i) = \sup_{\xi_i} \left\{ \alpha_i \xi_i - \ell(\xi_i) \right\}. \tag{21}$$

On note par ailleurs l'usage de l'identité

$$f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \ell(\xi_i) \Longrightarrow f^*(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \ell^*(\xi_i) \tag{22}$$

À présent, considérons le second terme de (20). Puisque l'expression est dérivable, on peut résoudre analytiquement q.

$$\nabla_q \left\{ \sum_{i=1}^{n} \alpha_i r_i q^T x_i + n\lambda\|q\|^2 \right\} = 0 \tag{23}$$

implique que

$$q = -\frac{1}{2n\lambda} \sum_{i=1}^{n} \alpha_i r_i x_i \tag{24}$$

à l'infimum.

En utilisant (24), on peut éliminer q de (20) pour obtenir

$$g(\alpha) = -\sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{2n\lambda} \sum_{i,j=1}^{n} \alpha_i \alpha_j r_i r_j x_i^T x_j + \frac{1}{4n\lambda} \sum_{i,j=1}^{n} \alpha_i \alpha_j r_i r_j x_i^T x_j \tag{25}$$

$$= -\sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{4n\lambda} (\alpha \circ r)^T K (\alpha \circ r). \tag{26}$$

Ainsi, sous sa forme duale, le problème (13) est équivalent à résoudre

$$\text{minimiser} \quad \sum_{i=1}^{n} \ell^*(\alpha_i) + \frac{1}{4n\lambda} (\alpha \circ r)^T K (\alpha \circ r). \tag{27}$$

On peut finalement définir $\tilde{\alpha}_i = \alpha_i / r_i$ pour obtenir le résultat annoncé plus haut.

Approche primale Soit $\kappa : \boldsymbol{X} \times \boldsymbol{X} \to \mathscr{R}$ un noyau semi-défini positif, $\boldsymbol{Q}$ l'espace de décision induit par κ et $K \in \mathscr{R}^{n \times n}$ la matrice de similarité. Le problème d'optimisation de portefeuille régularisé s'exprime alors par

$$\underset{q \in \boldsymbol{Q}}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i\, q(x_i)) - \lambda \|q\|^2. \tag{28}$$

Tel que mentionné, la dimension de $\boldsymbol{Q}$ est possiblement infinie, ce qui rend numériquement impossible la recherche d'une solution $q^\star$. Toutefois, le théorème de la représentation permet de rendre le problème résoluble.

Théorème 1 (Théorème de la représentation). *Toute solution $q^\star$ de* (28) *repose dans le sous-espace vectoriel engendré par l'ensemble des n fonctions $\{\phi_i\}$, où $\phi_i = \kappa(x_i, \cdot)$. Numériquement, il existe un vecteur $\alpha \in \mathscr{R}^n$ tel que,*

$$q^\star = \sum_{i=1}^{n} \alpha_i \phi_i = \alpha^T \phi. \tag{29}$$

Démonstration. Voir [MRT12], Théorème 5.4 pour une démonstration tenant compte d'un objectif régularisé général. La démonstration est dûe à [KW71]. □

Le théorème de la représentation permet donc de chercher une solution dans un espace à n dimensions, plutôt que la dimension possiblement infinie de $\boldsymbol{Q}$. En effet, puisque

$$q^\star = \sum_{i=1}^{n} \alpha_i \phi_i, \tag{30}$$

où $\alpha \in \mathscr{R}^n$, on peut donc restreindre le domaine d'optimisation à $\mathscr{R}^n$. L'objectif de (28) devient alors

$$n^{-1} \sum_{i=1}^{n} u(r_i \sum_{j=1}^{p} \alpha_j \phi_j(x_i)) - \lambda \langle q, q \rangle_{\boldsymbol{Q}}. \tag{31}$$

Le premier terme se réexprime comme

$$n^{-1} \sum_{i=1}^{n} u(r_i \, \alpha^T \phi(x_i)), \tag{32}$$

alors qu'en employant les propriétés de linéarité du produit intérieur, on transforme le second terme par

$$\langle q, q \rangle^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_i \alpha_j \langle \phi_i, \phi_j \rangle \tag{33}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_i \alpha_j \kappa(x_i, x_j) \tag{34}$$

$$= \alpha^T K \alpha. \tag{35}$$

De sorte que le problème général (28) peut se reformuler par

$$\underset{\alpha \in \mathscr{R}^n}{\text{maximiser}} \quad n^{-1} \sum_{i=1}^{n} u(r_i \alpha^T \phi(x_i)) - \lambda \, \alpha^T K \alpha \ . \tag{36}$$

Références

[KW71] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1) :82–95, 1971.

[MRT12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.