

CORRECTION DU TD 8 : MÉTHODES À NOYAUX

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, 13 NOVEMBRE 2015

Jean-Baptiste Alayrac
jean-baptiste.alayrac@inria.fr

1. EXEMPLES DE NOYAUX DÉFINIS POSITIFS

Dans cet exercice, \mathcal{X} est l'ensemble sur lequel sont définis nos noyaux. On rappelle qu'un noyau $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ est défini positif si :

$$\forall (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n, \forall (x_1, \dots, x_n) \in \mathcal{X}^n, \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \geq 0.$$

1) Opération sur les noyaux : Soient K et L deux noyaux définis positifs sur \mathcal{X} .

- (1) Montrons que $H := K + L$ est aussi un noyau défini positif. La symétrie de $K + L$ est évidente. Montrons alors le caractère défini positif du noyau de deux manière différentes :

- *En passant par la définition :* Soient $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ et $(x_1, \dots, x_n) \in \mathcal{X}^n$. On a alors :

$$\sum_{i,j} \alpha_i \alpha_j H(x_i, x_j) = \underbrace{\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)}_{\geq 0} + \underbrace{\sum_{i,j} \alpha_i \alpha_j L(x_i, x_j)}_{\geq 0} \geq 0.$$

- *Version noyau :* On sait qu'il existe un espace de Hilbert \mathcal{K} et \mathcal{L} tels qu'on ait $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}}$ et $L(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{L}}$. Définissons alors le vecteur $\theta(x)$ comme étant la concaténation des vecteurs $\phi(x)$ et $\psi(x)$. En dimension fini cela ne pose aucun problème. Nous pouvons juste supposer cela par la suite, cette preuve servant plus de donner une intuition sur comment se transforme le feature space en appliquant une opération plutôt que de donner une preuve réellement formelle. Dans ce cas on a alors que :

$$H(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}} + \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{L}} = \langle \theta(x_i), \theta(x_j) \rangle.$$

Par suite le noyau H est bien défini positif. En effet ici nous avons réussi à réécrire le noyau comme un produit scalaire. Cette "astuce" sera souvent utilisée donc rappelons simplement que :

$$\sum_{i,j} \alpha_i \alpha_j \langle \theta(x_i), \theta(x_j) \rangle = \left\| \sum_i \alpha_i \theta(x_i) \right\|^2 \geq 0,$$

ce qui justifie bien le caractère défini positif.

- (2) Montrons que $H := KL$ est aussi un noyau défini positif.

- *En passant par la définition* : Soient $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ et $(x_1, \dots, x_n) \in \mathcal{X}^n$. Notons M_K et M_L les matrices de Gramm associés respectivement à K et L . Montrons alors que $M_H = M_K \odot M_L$ (où \odot est le produit d'Hadamard ou produit terme à terme de deux matrices) est semi défini positive. Comme M_K est une matrice symétrique semi défini positive on sait qu'on peut la décomposer sous la forme $M_K = \sum_i \lambda_i u_i u_i^T$, où tous les λ_i sont positifs. On a alors :

$$M_H = \sum_i \lambda_i u_i u_i^T \odot M_L$$

Or on a que pour un vecteur u quelconque :

$$\sum_{ij} \alpha_i \alpha_j (u u^T \odot M_L)_{ij} = \sum_{ij} \alpha_i \alpha_j (M_L)_{ij} u_i u_j = (\alpha \odot u)^T M_L (\alpha \odot u) \geq 0$$

Donc par somme de termes positifs on aura bien $\alpha^T M_H \alpha \geq 0$.

- *Version noyau* : On sait qu'il existe un espace de Hilbert \mathcal{K} et \mathcal{L} tels qu'on ait $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}}$ et $L(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{L}}$. Définissons alors le vecteur $\theta(x)$ comme étant la concaténation des vecteurs $\phi(x)$ et $\psi(x)$. Même remarque que plus haut en ce qui concerne la dimension. On a alors que :

$$\begin{aligned} H(x_i, x_j) &= \phi(x_i)^T \phi(x_j) \psi(x_j)^T \psi(x_i) \\ &= \text{Tr}(\phi(x_i)^T \phi(x_j) \psi(x_j)^T \psi(x_i)) \\ &= \text{Tr}(\phi(x_j) \psi(x_j)^T \psi(x_j)^T \phi(x_i)^T) \\ &= \langle \phi(x_j) \psi(x_j)^T, \phi(x_i) \psi(x_i)^T \rangle, \end{aligned}$$

où le produit scalaire du dessus est le produit scalaire usuel sur les matrices. Par suite H est bien un noyau défini positif.

2) Minimum : $\mathcal{X} = \mathbb{R}^+$, $K(x, y) = \min(x, y)$. Soient $(x_1, \dots, x_n) \in \mathbb{R}_+^n$. Sans perte de généralités on peut supposer que les x_i sont ordonnées. Alors en écrivant la matrice M_K on voit qu'on peut triangulariser la matrice de manière simple en opérant sur ses lignes. Les valeurs propres de la matrice se lisent alors sur la diagonale et ont pour valeurs : $(x_1, x_2 - x_1, \dots, x_n - x_{n-1})$. Ces valeurs propres sont toutes positives ou nulles, par suite la matrice M_K est semi définie positive et donc le noyau K est bien défini positif.

Remarque : une preuve plus élégante consiste à dire que ce noyau est en fait la covariance d'un mouvement brownien...

3) Chi-2 : $\mathcal{X} = \mathbb{R}_*^+$, $K(x, y) = 2 \frac{xy}{x+y}$.

xy est clairement un noyau défini positif (noyau linéaire). Pour $\frac{1}{x+y}$ on le réécrit comme un produit scalaire :

$$\frac{1}{x+y} = \int_0^1 t^{x-\frac{1}{2}} t^{y-\frac{1}{2}} dt = \langle \phi(x), \phi(y) \rangle_{L_2}.$$

On peut donc conclure en utilisant la propriété sur les produits de noyaux démontrée plus haut.

4) Sur des ensembles : $\mathcal{X} = \mathcal{P}(\mathcal{A})$ avec \mathcal{A} un ensemble de cardinal fini. $K(A, B) = \frac{|A \cap B|}{|A \cup B|}$. Notons n le cardinal de \mathcal{A} . En notant $\phi(A) \in 0, 1^n$ le vecteur indicateur de l'ensemble A , on a d'une part :

$$|A \cap B| = \phi(A)^T \phi(B)$$

D'autre part on a (en notant A^c le complémentaire de A) :

$$\begin{aligned} \frac{1}{|A \cup B|} &= \frac{1}{n - |A^c \cap B^c|} \\ &= \frac{1}{n(1 - \frac{|A^c \cap B^c|}{n})} \\ &= \frac{1}{n(1 - \frac{\phi(A^c)^T \phi(B^c)}{n})} \\ &= \frac{1}{n} \sum_{i=0}^{\infty} \left(\frac{\phi(A^c)^T \phi(B^c)}{n} \right)^i \end{aligned}$$

Or chaque $(\frac{\phi(A^c)^T \phi(B^c)}{n})^i$ définit un noyau défini positif entre A et B (produit de noyaux définis à partir de produit scalaire). On peut aussi prouver (faire en exercice) que la limite de noyaux défini positifs qui converge point par point est aussi un noyau défini positif. On peut alors conclure et dire que $\frac{1}{|A \cup B|}$ est un noyau défini positif.

Par produit K est donc un noyau.

5) Bonus : $\mathcal{X} = \mathbb{N}$, $K(n, m) = \text{PGCD}(n, m)$.

Ecrire :

$$\text{PGCD}(n, m) = \prod_{p_i} p_i^{\min(\phi_i(m), \phi_i(n))},$$

où les p_i sont les nombres premiers et où $\phi_i(m)$ donne la valuation de p_i dans la décomposition en facteur premier de m ...

2. MANIPULATION DE LA DISTANCE DANS LE FEATURE SPACE

6) a) Soient $(x, y) \in \mathcal{X}$. Soit K un noyau défini positif sur \mathcal{X} . On rappelle qu'il existe un espace de Hilbert \mathcal{F} pour lequel on a $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$. On a alors :

$$\|\phi(x) - \phi(y)\|_{\mathcal{F}}^2 = K(x, x) - 2K(x, y) + K(y, y)$$

b) Pour le noyau Chi-2 on retrouve la distance Chi-2 :

$$\|\phi(x) - \phi(y)\|_{\mathcal{F}}^2 = \frac{(x - y)^2}{x + y}.$$

7) Distance à la moyenne dans le feature space. On considère ici des points $(x_1, \dots, x_n) \in \mathcal{X}^n$ et des réponses binaires associées $(y_1, \dots, y_n) \in \{-1, 1\}^n$. Soit un noyau K défini positif sur \mathcal{X} . On se propose d'étudier une règle de classification très simple qui va simplement décider en fonction des distances aux centroïdes respectifs de chaque classe.

a) Soit $x \in \mathcal{X}$. La distance entre $\phi(x)$ et $\frac{1}{n_+} \sum_{i, y_i=1} \phi(x_i)$ uniquement en fonction de K (où n_+ est le nombre de y_i tels que $y_i = 1$) vaut alors :

$$\|\phi(x) - \frac{1}{n_+} \sum_{i, y_i=1} \phi(x_i)\|^2 = K(x, x) + \frac{1}{n_+^2} \sum_{i, y_i=1} \sum_{j, y_j=1} K(x_i, x_j) - \frac{2}{n_+} \sum_{i, y_i=1} K(x, x_i)$$

b) Une règle de classification simple pour le vecteur x en fonction des données $(x_i)_{i=1, \dots, n}$ et $(y_i)_{i=1, \dots, n}$ et du noyau K est simplement :

$$(1) \quad y_i = \begin{cases} 1 & \|\phi(x) - \frac{1}{n_+} \sum_{i, y_i=1} \phi(x_i)\|^2 \leq \|\phi(x) - \frac{1}{n_-} \sum_{i, y_i=-1} \phi(x_i)\|^2 \\ -1 & \text{sinon} \end{cases}$$

c) Supposons maintenant que $\frac{1}{n} \sum_i \phi(x_i) = 0$ (donnée centrée dans \mathcal{F}) ainsi que $\sum_i y_i = 0$ (autant de points positifs que négatifs). On a alors :

$$\|\frac{1}{n_+} \sum_{i, y_i=1} \phi(x_i)\|^2 = \|\frac{1}{n_-} \sum_{i, y_i=-1} \phi(x_i)\|^2$$

On peut alors simplifier la règle de classification de la sorte :

$$\|\phi(x) - \frac{1}{n_+} \sum_{i, y_i=1} \phi(x_i)\|^2 \leq \|\phi(x) - \frac{1}{n_-} \sum_{i, y_i=-1} \phi(x_i)\|^2 \Leftrightarrow \sum_{i, y_i=1} K(x, x_i) \geq \sum_{i, y_i=-1} K(x, x_i)$$

d) L'application est directe et sert plus ici à remarquer que cette règle de classification à priori triviale permet de retomber sur une méthode assez élaborée vu dans le cadre du moyennage local.

Conclusion : Tout ce qu'on a utilisé ici se résume à "l'astuce du noyau" (qui permet de réécrire des quantités uniquement en fonction de K). Cette astuce d'apparence triviale débouche sur d'importantes applications. Elle permet notamment d'obtenir des versions non linéaires d'algorithmes linéaires que nous avons vu jusqu'à présent (en remplaçant le produit scalaire usuel par le noyau gaussien par exemple). Ceci peut alors être utilisé pour appliquer des méthodes classiques à des données non vectorielles (page web, séquence ADN...) mais aussi pour enrichir des méthodes qui existent déjà sur des données vectorielles.