Classification non supervisée et données fonctionnelles

Aurélie FISCHER

Directeur de M2 : Pascal MASSART
Directeur de thèse : Gérard BIAU

Octobre 2008

Table des matières

1 Données fonctionnelles

On s'intéresse à des données de dimension infinie ou, ce qui revient au même, de très grande dimension.

Une donnée fonctionnelle est en réalité constituée d'un certain nombre de valeurs discrètes que l'on a mesurées, enregistrées, mais dont l'ensemble reflète une variation régulière. Comme en théorie, on pourrait obtenir une grande quantité de points, aussi rapprochés que l'on veut, on voit que l'on obtient une courbe et que l'on peut traiter la donnée comme une fonction.

Remarque. Considérer comme des fonctions des données qui après tout sont discrètes permet justement de faire intervenir la régularité de ces fonctions. Par exemple, pour un phénomène donné, grâce à l'utilisation de données fonctionnelles, on pourra étudier aussi sa vitesse ou son accélération.

2 Classification non supervisée

Il s'agit de classer les données en k groupes de telle sorte qu'elles soient très semblables à l'intérieur d'un groupe, les différents groupes étant aussi séparés que possible. Le nombre de groupes k n'est pas forcément fixé d'avance.

On se place dans un espace de Hilbert $\mathcal{H}$ séparable. On note $\langle \cdot, \cdot \rangle$ le produit scalaire dans $\mathcal{H}$ et $\| \cdot \|$ la norme associée. Soit X une variable aléatoire de loi μ à valeurs dans $\mathcal{H}$ et $X_1, \dots, X_n$ un n-échantillon de loi μ. Supposons que k est fixé et $\mathbb{E}\|X\|^2 < \infty$. On s'intéresse au problème de classification non supervisée pour les observations $X_1, \dots, X_n$.

2.1 Méthode des centres mobiles

Une méthode de classification consiste à utiliser l'algorithme des centres mobiles. Afin de ranger les données dans k classes, on minimise le risque de classification empirique

$$W(\mathbf{c}, \mu_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|X_i - c_j\|^2$$

sur tous les choix possibles pour les k centres $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{H}^k$. μ_n est la loi empirique des données, définie par

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}$$

pour tout borélien $A \subset \mathcal{H}$.

Définition 2.1 (Ensemble de Voronoi). *On appelle ensemble de Voronoi associé au centre c_j, le polyèdre convexe de tous les points de $\mathcal{H}$ qui sont plus proches de c_j que de tout autre centre. On le note S_j. Un point à égale distance de plusieurs centres est affecté arbitrairement à l'un d'eux.*

Au cours de l'étape $r+1$ de l'algorithme, chaque X_i est affecté au centre $c_{nj}^{(r)}$ le plus proche, puis le "nouveau centre empirique" $c_{nj}^{(r+1)}$ est calculé en faisant la moyenne des X_i tombés dans $S_j^{(r)}$.

2.2 Risque de classification

Lorsque l'algorithme donne un ensemble de centres $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{H}^k$, l'erreur en moyenne quadratique ou risque de classification

$$W(\mathbf{c}, \mu) = \int \min_{j=1,\ldots,k} \|x - c_j\|^2 d\mu(x)$$

permet d'en mesurer la performance. On désigne par

$$W^*(\mu) = \inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu)$$

le risque optimal.

Définition 2.2 (δ_n–minimiseur). *Soit $\delta_n \geq 0$. Un ensemble de centres $\mathbf{c}_n = (c_{n1}, \ldots, c_{nk}) \in \mathcal{H}^k$ qui vérifie*

$$W(\mathbf{c}_n, \mu_n) \leq W^*(\mu_n) + \delta_n$$

est appelé un δ_n–minimiseur du risque de classification empirique. Lorsque $\delta_n = 0$, on dit que $\mathbf{c}_n$ est un minimiseur du risque de classification empirique.

On peut obtenir la borne suivante :

Théorème 2.1. *Soit $\mathbb{E}\|X\|^2 < \infty$. Pour tout $x > 0$, il existe des constantes $C(\mu) > 0$ et $N_0(\mu, k, x) > 0$ telles que, pour tout $n \geq N_0$, avec probabilité au moins $1 - 2e^{-x}$,*

$$W(\mathbf{c}_n, \mu) - W^*(\mu) \leq C(\mu) \frac{k + \sqrt{x}}{\sqrt{n}} + \delta_n$$

2.3 Projection aléatoire baséee sur la méthode de Johnson-Lindenstrauss

2.3.1 Projection aléatoire

Soit $s \in \mathbb{N}^*$, et soient $(N_{1\alpha})_{\alpha\geq 1}, \ldots, (N_{s\alpha})_{\alpha\geq 1}$ s suites indépendantes de variables aléatoires gaussiennes de loi $\mathcal{N}\left(0, \frac{1}{s}\right)$. $\forall D \geq 1$, soit

$$\overline{X}_{ij}^D = \sum_{\alpha=1}^{D} N_{j\alpha} X_{i\alpha}, \quad j \in \{1, \ldots, s\}$$

c'est-à-dire, sous forme matricielle,

$$\begin{pmatrix} \overline{X}_{i1}^D \\ \vdots \\ \overline{X}_{is}^D \end{pmatrix} = \begin{pmatrix} N_{11} & \ldots & N_{1D} \\ \vdots & \vdots & \vdots \\ N_{s1} & \ldots & N_{sD} \end{pmatrix} \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iD} \end{pmatrix}$$

Sachant les données $X_1, \ldots, X_n$, pour tous i, j fixés, comme $\overline{X}_{ij}^D$ est une somme de variables aléatoires centrées indépendantes et $\overline{X}_{ij}^D \in \mathbb{L}^1$, la suite $(\overline{X}_{ij}^D)_{D\geq 1}$ est une martingale par rapport à la filtration canonique. Or,

$$\mathbb{E}(\overline{X}_{ij}^D)^2 \big| X_1, \ldots, X_n = \sum_{\alpha=1}^{D} \frac{(X_{i\alpha})^2}{s} \leq \frac{\|\mathbf{X}_i\|^2}{s}$$

(espérance par rapport aux variables gaussiennes) Donc, la martingale $(\overline{X}_{ij}^D)_{D\geq 1}$ est bornée dans $\mathbb{L}^2$. On en déduit qu'elle converge p.s. et dans $\mathbb{L}^2$ et qu'il existe une variable aléatoire $X_{ij}^* \in \mathbb{L}^2$ telle que $|\overline{X}_{ij}^D| \leq X_{ij}^*$. Soit $\overline{X_{ij}}$ la limite de $(\overline{X}_{ij}^D)_{D\geq 1}$. Par convergence dominée, on a

$$\lim_{D\to\infty} \mathbb{E}(\overline{X}_{ij}^D)^2 \big| X_1, \ldots, X_n = \mathbb{E}(\overline{X}_{ij})^2 \big| X_1, \ldots, X_n = \frac{\|\mathbf{X}_i\|^2}{s}$$

A chaque donnée X_i, on peut ainsi associer un vecteur $\overline{\mathbf{X}}_i = (\overline{X}_{i1}, \ldots, \overline{X}_{is}) \in \mathbb{R}^s$, qui est sa "projection aléatoire sur un sev de dimension finie s". Chaque composante de $\overline{\mathbf{X}}_i$ est une variable aléatoire gaussienne de loi $\mathcal{N}\left(0, \frac{\|\mathbf{X}_i\|^2}{s}\right)$. Donc,

$$\mathbb{E}\|\overline{\mathbf{X}}_i\|^2 \big| X_1, \ldots, X_n = \|\mathbf{X}_i\|^2$$

et

$$\frac{s\|\overline{\mathbf{X}}_i\|^2}{\|\mathbf{X}_i\|^2} \sim \chi^2(s)$$

De même,

$$\forall i \neq i' \in \{1, \dots, n\}, \quad \mathbb{E}\|\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_{i'}\|^2 \big| X_1, \dots, X_n = \|\mathbf{X}_i - \mathbf{X}_{i'}\|^2$$

et

$$\frac{s\|\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_{i'}\|^2}{\|\mathbf{X}_i - \mathbf{X}_{i'}\|^2} \sim \chi^2(s)$$

L'inégalité de Chernoff pour la loi χ^2 entraîne

$$\mathbb{P}\left\{\frac{\|\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_{i'}\|^2}{\|\mathbf{X}_i - \mathbf{X}_{i'}\|^2} - 1 > \varepsilon \Big| X_1, \dots, X_n\right\} \leq \exp\left[\frac{s}{2}(-\varepsilon + \ln(1+\varepsilon))\right]$$

ainsi que

$$\mathbb{P}\left\{\frac{\|\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_{i'}\|^2}{\|\mathbf{X}_i - \mathbf{X}_{i'}\|^2} - 1 < -\varepsilon \Big| X_1, \dots, X_n\right\} \leq \exp\left[\frac{s}{2}(\varepsilon + \ln(1-\varepsilon))\right]$$

Ces considérations entraînent le lemme de Johnson-Lindenstrauss :

2.3.2 Lemme de Johnson-Lindenstrauss

Le lemme établit que, pour $\varepsilon \in \ ]0,1[$, tout ensemble de n points dans un espace de Hilbert séparable peut se plonger dans un espace euclidien de dimension $O(\ln \frac{n}{\varepsilon^2})$ en modifiant peu les distances, qui au pire sont multipliées par un facteur $1 \pm \varepsilon$.

Lemme 2.1 (Johnson-Lindenstrauss). *Soit $\mathcal{H}$ un espace de Hilbert séparable. $\forall \varepsilon, \delta \in \ ]0,1[, n \in \mathbb{N}^*$, soit $s \in \mathbb{N}^*$ tel que $s \geq \frac{4}{\varepsilon^2/2 - \varepsilon^3/3} \ln \frac{n}{\sqrt{\delta}}$ Soit $f : \mathcal{H} \to \mathbb{R}^s$ une projection aléatoire (construite comme dans le paragraphe précédent). Alors, pour tout ensemble $\mathcal{D}$ de n points de $\mathcal{H}$, avec probabilité au moins $1 - \delta$, on a*

$$\forall (u, v) \in \mathcal{D}^2, (1-\varepsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1+\varepsilon)\|u - v\|^2$$

2.3.3 Algorithme

On applique l'algorithme des centres mobiles aux projections $\overline{\mathbf{X}}_1, \dots, \overline{\mathbf{X}}_n \in \mathbb{R}^s$: on obtient un ensemble $\overline{\mathbf{c}}_n$ de k centres $\overline{c}_{n1}, \dots, \overline{c}_{nk} \in \mathbb{R}^s$ qui minimisent le risque de classification empirique correspondant aux projections. Soient $\overline{S}_{n1}, \dots, \overline{S}_{nk} \subset \mathbb{R}^s$ les ensembles de Voronoi associés. On introduit alors $\hat{\mathbf{c}}_n$, l'ensemble des centres dans $\mathcal{H}$ que l'on définit comme suit :

$$\forall j \in \{1, \dots, k\}, \quad \hat{c}_{nj} = \frac{\sum_{i=1}^n X_i \mathbb{1}_{\{\overline{\mathbf{X}}_i \in \overline{S}_{nj}\}}}{\sum_{i=1}^n \mathbb{1}_{\{\overline{\mathbf{X}}_i \in \overline{S}_{nj}\}}}$$

A partir de ces centres, on déduit ensuite la partition de Voronoi de $\mathcal{H}$ en k classes.

Théorème 2.2. *$\forall \varepsilon, \delta \in \ ]0,1[$, soient l'entier s et la projection aléatoire f comme dans le lemme de Johnson-Lindenstrauss. On a*

$$W(\hat{\mathbf{c}}_n, \mu_n) \leq \frac{1+\varepsilon}{1-\varepsilon} W(\mathbf{c}_n, \mu_n)$$

avec probabilité au moins $1-\delta$.

Démonstration. On pose

$$\overline{N}_j = \sum_{i=1}^{n} \mathbb{1}_{\{\overline{\mathbf{X}}_i \in \overline{S}_{nj}\}}$$

et

$$N_j = \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{X}_i \in S_{nj}\}}$$

$$\begin{aligned} W(\overline{\mathbf{c}}_n, \mu_n) &= \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\overline{\mathbf{X}}_i - \overline{c}_{nj}\|^2 \\ &= \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \|\overline{\mathbf{X}}_i - \overline{c}_{nj}\|^2 \mathbb{1}_{\{\overline{\mathbf{X}}_i \in \overline{S}_{nj}\}} \\ &= \sum_{j=1}^{k} \frac{1}{2n\overline{N}_j} \sum_{i_1,i_2=1}^{n} \|\overline{\mathbf{X}}_{i_1} - \overline{\mathbf{X}}_{i_2}\|^2 \mathbb{1}_{\{(\overline{\mathbf{X}}_{i_1}, \overline{\mathbf{X}}_{i_2}) \in \overline{S}_{nj}^2\}} \end{aligned}$$

En raison de l'optimalité de la méthode des centres mobiles (Linder [10], lemme 1), on a, considérant l'algorithme appliqué aux projections $\overline{\mathbf{X}}_i$, $i \in \{1, \ldots, n\}$,

$$W(\overline{\mathbf{c}}_n, \mu_n) \leq \sum_{j=1}^{k} \frac{1}{2nN_j} \sum_{i_1,i_2=1}^{n} \|\overline{\mathbf{X}}_{i_1} - \overline{\mathbf{X}}_{i_2}\|^2 \mathbb{1}_{\{(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) \in S_{nj}^2\}}$$

D'après le lemme de Johnson-Lindenstrauss,

$$W(\overline{\mathbf{c}}_n, \mu_n) \leq (1+\varepsilon) \sum_{j=1}^{k} \frac{1}{2nN_j} \sum_{i_1,i_2=1}^{n} \|\mathbf{X}_{i_1} - \mathbf{X}_{i_2}\|^2 \mathbb{1}_{\{(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) \in S_{nj}^2\}}$$

c'est-à-dire

$$W(\overline{\mathbf{c}}_n, \mu_n) \leq (1+\varepsilon) W(\mathbf{c}_n, \mu_n) \tag{1}$$

De même, le lemme de Johnson-Lindenstrauss implique

$$(1-\varepsilon) \sum_{j=1}^{k} \frac{1}{2n\overline{N}_j} \sum_{i_1,i_2=1}^{n} \|\mathbf{X}_{i_1} - \mathbf{X}_{i_2}\|^2 \mathbb{1}_{\{(\overline{\mathbf{X}}_{i_1}, \overline{\mathbf{X}}_{i_2}) \in \overline{S}_{nj}^2\}} \leq W(\overline{\mathbf{c}}_n, \mu_n)$$

donc

$$(1-\varepsilon) W(\hat{\mathbf{c}}_n, \mu_n) \leq W(\overline{\mathbf{c}}_n, \mu_n) \tag{2}$$

En combinant les inégalités (1) et (2), on obtient le résultat cherché. □

2.4 Un complément possible

Au lieu de fixer à l'avance le nombre de groupes k, on pourrait s'intéresser à un critère de la forme

$$\inf_k \left(\frac{1}{n} \sum \min_{j=1,\dots,k} \|X_i - c_j\|^2 + \text{pen}(k) \right)$$

avec pen(k) une pénalité convenable pour éviter de prendre k trop grand. Le choix $k = n$ minimise en effet le risque empirique, mais ne classe pas les données !

3 Quantification

Soit $(E, \|\cdot\|)$ un espace de Banach réflexif et séparable. Soit X une variable aléatoire de loi μ à valeurs dans E, telle que $\mathbb{E}(\|X\|) < \infty$.

Définition 3.1 (Quantificateur). *Soit $k \geq 1$ un entier. Un quantificateur de k points est une application borélienne $q : \mathbb{E} \to \mathcal{C}$ où $\mathcal{C} = \{y_1, \dots, y_k\}$ est un ensemble de k éléments de E appelé table de codage.*

Remarque. $y_1, \dots, y_k$ seront également appelés les centres associés au quantificateur q.

Pour tout $x \in E$, on représente x par un unique $\hat{x} = q(x) \in \mathcal{C}$. Pour mesurer l'erreur que l'on fait en représentant X par $q(X)$, on s'intéresse à $d(X, q(X))$, où d est une fonction mesurable vérifiant $\forall (x, y) \in E^2$ $d(x, y) \geq 0$, appelée mesure de distorsion.

Définition 3.2. *La distorsion de q quantifiant X est la quantité*

$$W(\mu, q) = \mathbb{E}d(X, q(X)) = \int_E d(x, q(x))\mu(dx)$$

On cherche le quantificateur de k points qui minimise cette distorsion.

Définition 3.3 (Quantificateur optimal). *Soit Q_k l'ensemble de tous les quantificateurs de k-points, et soit*

$$W_k^*(\mu) = \inf_{q \in Q_k} W(\mu, q)$$

Un quantificateur $q^* \in Q_k$ est dit optimal si $W(\mu, q^*) = W_k^*(\mu)$.

Remarque. Tout quantificateur de k points est déterminé par sa table de codage $\{y_i\}_{i=1}^k$ et une partition de E en cellules $S_i = \{x : q(x) = y_i\}$, $i = 1, \dots, k$: on a la relation

$$q(x) = y_i \Leftrightarrow x \in S_i$$

On peut donc définir un quantificateur par sa table de codage et sa partition en cellules.

4 Divergence de Bregman

Une distance généralisée est une fonction réelle positive de deux variables $d(x, y)$ utilisée pour mesurer la distance entre x et y en un certain sens généralisé. Une distance généralisée ne vérifie pas toujours l'inégalité triangulaire, et elle n'est pas forcément symétrique : on peut avoir $d(x, y) \neq d(y, x)$. On va utiliser comme mesure de distorsion une famille de distances généralisées qui a été introduite par Bregman en 1967 dans [5].

4.1 Définition et exemples dans $\mathbb{R}^d$

Définition 4.1. *On appelle intérieur relatif d'un convexe non vide $\mathcal{C}$, noté $ir(\mathcal{C})$, l'intérieur de $\mathcal{C}$ relativement au sous-espace affine engendré par $\mathcal{C}$.*

Remarque. Alors que l'intérieur d'un convexe est souvent vide, ce n'est pas le cas de son intérieur relatif, d'où l'intérêt d'introduire cette notion : en dimension finie, l'intérieur relatif d'un convexe non vide $\mathcal{C}$ est non vide, (et a la même dimension que $\mathcal{C}$).

Définition 4.2. *Soient $\mathcal{C} \subset \mathbb{R}^d$ un convexe et $\phi : \mathcal{C} \mapsto \mathbb{R}$ une fonction strictement convexe, différentiable sur $ir(\mathcal{C})$. La divergence de Bregman associée $d_\phi : \mathcal{C} \times ir(\mathcal{C}) \mapsto [0, \infty[$ est définie par*

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle$$

La distance euclidienne et d'autres distances usuelles sont des cas particuliers de divergence de Bregman.

Exemple 4.1. Soient $\mathcal{C} = \mathbb{R}^d$ et ϕ définie par $\phi(x) = \|x\|^2$. ϕ est strictement convexe et différentiable sur $\mathbb{R}^d$.

$$\begin{aligned} d_\phi(x, y) &= \|x\|^2 - \|y\|^2 - \langle x - y, \nabla\phi(y)\rangle \\ &= \|x\|^2 - \|y\|^2 - \langle x - y, 2y\rangle \\ &= \|x - y\|^2 \end{aligned}$$

On obtient le carré de la distance euclidienne.

Exemple 4.2. Soit $\mathcal{C}$ le simplexe de dimension d, p une mesure de probabilité discrète telle que $\sum_{j=1}^d p_j = 1$. La fonction ϕ définie par $\phi(p) = \sum_{j=1}^d p_j \ln p_j$ est convexe. La divergence de Bregman obtenue avec ce choix de ϕ est la distance de

Kullback-Leibler.

$$\begin{aligned} d_\phi(p,q) &= \sum_{j=1}^{d} p_j \ln p_j - \sum_{j=1}^{d} q_j \ln q_j - \langle p-q, \nabla\phi(q)\rangle \\ &= \sum_{j=1}^{d} p_j \ln p_j - \sum_{j=1}^{d} q_j \ln q_j - \sum_{j=1}^{d} (p_j - q_j)(\ln q_j + 1) \\ &= \sum_{j=1}^{d} p_j \ln \frac{p_j}{q_j} \end{aligned}$$

4.2 Cas fonctionnel

Définition 4.3 (Divergence de Bregman fonctionnelle). *Soit $\phi : E \to \mathbb{R}$ de classe C^2 strictement convexe. La divergence de Bregman associée est définie par*

$$d_\phi(x,y) = \phi(x) - \phi(y) - D_y\phi(x-y)$$

avec $D_y\phi$ la différentielle de ϕ en y.

Voici quelques exemples de divergences de Bregman fonctionnelles, introduits dans [6].

Exemple 4.3. Soit $E = L^2(m)$ muni de la norme $\|\cdot\|_{L^2(m)}$ et $\phi : x \mapsto \int x^2 dm$. On a

$$\begin{aligned} \phi(y+h) - \phi(h) &= \int (y+h)^2 dm - \int y^2 dm \\ &= 2\int yh dm + \int h^2 dm \end{aligned}$$

Comme $\lim_{\|h\|\to 0} \frac{\int h^2 dm}{\|h\|} = \lim_{\|h\|\to 0} \|h\| = 0$ et que $h \mapsto 2\int yh dm$ est une application linéaire continue, on a $D_y\phi(h) = 2\int yh dm$.

$$\begin{aligned} D_y^2\phi(k,h) &= D_{y+k}\phi(h) - D_y\phi(h) \\ &= 2\int (y+k)h dm - 2\int yh dm \\ &= 2\int kh dm \end{aligned}$$

D'où $D_y^2\phi(h,h) = 2\|h\|^2$, ce qui montre que la fonction ϕ est strictement convexe.

Finalement,

$$\begin{aligned} d_\phi(x,y) &= \int x^2 dm - \int y^2 dm - 2\int y(x-y)dm \\ &= \int (x-y)^2 dm \\ &= \|x-y\|^2_{L^2(m)} \end{aligned}$$

Exemple 4.4. Soit $E = \{x \in L^1(m), x \geq 0\}$ muni de la norme $\|\cdot\|_{L^1(m)}$ et $\phi : x \mapsto \left(\int x dm\right)^2$. On a

$$\begin{aligned} \phi(y+h) - \phi(h) &= \left(\int y dm + \int h dm\right)^2 - \left(\int y dm\right)^2 \\ &= 2\int y dm \int h dm + \left(\int h dm\right)^2 \end{aligned}$$

L'application $h \mapsto 2\int y dm \int h dm$ est linéaire continue et on a $0 \leq \frac{\left(\int h dm\right)^2}{\|h\|} \leq \frac{\|h\|^2}{\|h\|} = \|h\|$, donc $D_y\phi(h) = 2\int y dm \int h dm$

$$\begin{aligned} D^2_y\phi(k,h) &= D_{y+k}\phi(h) - D_y\phi(h) \\ &= 2\int (y+k)dm \int h dm - 2\int y dm \int h dm \\ &= 2\int k dm \int h dm \end{aligned}$$

Donc $D^2_y\phi(h,h) = 2\|h\|^2$, et ϕ est strictement convexe. La divergence de Bregman associée est

$$\begin{aligned} d_\phi(x,y) &= \left(\int x dm\right)^2 - \left(\int y dm\right)^2 - 2\int y dm \int (x-y)dm \\ &= \left(\int x dm\right)^2 + \left(\int y dm\right)^2 - 2\int y dm \int x dm \\ &= \left(\int (x-y)dm\right)^2 \end{aligned}$$

5 Quantificateur des plus proches voisins

Définition 5.1 (Partition de Voronoi). *Etant donné une table de codage* $\mathcal{C} = \{y_i\}_{i=1}^k$*, une partition* $\{S_i\}_{i=1}^k$ *vérifiant*

$$S_1 = \{x \in E, d_\phi(x, y_1) \leq d_\phi(x, y_j), j = 1 \ldots, k\}$$

et pour $i = 2, \ldots, k$,

$$S_i = \{x \in E, d_\phi(x, y_i) \leq d_\phi(x, y_j), j = 1, \ldots, k\} \backslash \bigcup_{\ell=1}^{i-1} S_\ell$$

est appelée partition de Voronoi.

Remarque. Retirer $\bigcup_{\ell=1}^{i-1} S_\ell$ est une manière d'éviter les ambiguités au bord des cellules : si $d_\phi(x, y_i) = d_\phi(x, y_j)$, on affecte x à la cellule dont l'indice est le plus petit.

Définition 5.2 (Quantificateur des plus proches voisins). *Un quantificateur de table de codage* $\mathcal{C} = \{y_i\}_{i=1}^k$ *qui admet comme partition la partition de Voronoi associée à* $\mathcal{C}$ *est appelé quantificateur des plus proches voisins.*

Lemme 5.1 (Meilleure partition). *Soient* q *un quantificateur de* k *points de table de codage* $\{y_i\}_{i=1}^k$ *et* q' *le quantificateur des plus proches voisins ayant même table de codage. Alors, on a*

$$W(\mu, q') \leq W(\mu, q)$$

Démonstration. Par définition de q', on a

$$d_\phi(x, q'(x)) = \min_{y \in \mathcal{C}} d_\phi(x, y)$$

Donc

$$\begin{aligned} W(\mu, q) &= \mathbb{E} d_\phi(X, q(X)) \\ &= \int_E d_\phi(x, q(x)) \mu(dx) \\ &= \sum_{j=1}^k \int_{S_j} d_\phi(x, y_j) \mu(dx) \\ &\geq \sum_{j=1}^k \int_{S_j} \min_{y \in \mathcal{C}} d_\phi(x, y) \mu(dx) \\ &= \mathbb{E} d_\phi(X, q'(X)) \\ &= W(\mu, q') \end{aligned}$$

□

Remarque. Cela nous montre que s'il existe un quantificateur optimal, c'est nécessairement un quantificateur des plus proches voisins. Donc il suffit de considérer les quantificateurs de ce type, et de trouver la table de codage optimale :

$$W_k^*(\mu) = \inf_{\mathcal{C} \subset E : |\mathcal{C}| = k} \mathbb{E} \min_{y \in \mathcal{C}} d_\phi(X, y)$$

Comparons à présent les quantificateurs de même partition, pour voir quelle est la meilleure table de codage. Pour cela, il faut montrer l'existence de

$$\arg\min_{y\in E} \mathbb{E}d_\phi(X,y)$$

Rappelons tout d'abord quelques notions d'optimisation, que l'on retrouvera notamment dans [7].

Définition 5.3 (Coercivité). *Une fonctionnelle $L : E \times E \to \mathbb{R}$ est dite coercive s'il existe $c > 0$ tel que $\forall x \in E, L(x,x) \geq c\|x\|^2$.*

Lemme 5.2 (Conditions d'existence d'un extremum). *Soit $L : E \to \mathbb{R}$ de classe C^2.*

- *Si L a un extremum en $y \in E$, alors $D_y L \equiv 0$.*
- *Soit $y \in E$. Si $D_y L \equiv 0$ et $D_y^2 L(\cdot,\cdot)$ est coercive, alors L admet un minimum en y.*

Le théorème suivant est dû à Frigyik et al. ([6]), et a été démontré dans le cas fini-dimensionnel par Banerjee et al. dans [1].

Théorème 5.1. *On suppose que $\phi : E \to \mathbb{R}$ est de classe C^3, et que $\forall y \in E$, la différentielle seconde $D_y^2\phi(\cdot,\cdot)$ est coercive. Alors $J : y \mapsto \mathbb{E}d_\phi(X,y)$ atteint son minimum sur E en $y = \mathbb{E}[X]$.*

Démonstration. Le développement de Taylor de ϕ nous donne :

$$\phi(y+h) = \phi(y) + D_y\phi(h) + \frac{1}{2}D_y^2\phi(h,h) + o(\|h\|^2) \tag{3}$$

$$\phi(y) = \phi(y+h) - D_{y+h}\phi(h) + \frac{1}{2}D_{y+h}^2\phi(h,h) + o(\|h\|^2) \tag{4}$$

En additionnant (3) et (4), on obtient :

$$0 = D_y\phi(h) - D_{y+h}\phi(h) + \frac{1}{2}D_y^2\phi(h,h) + \frac{1}{2}D_{y+h}^2\phi(h,h) + o(\|h\|^2) \tag{5}$$

Comme on a $|D^2\phi_{y+h}(h,h) - D_y^2\phi(h,h)| \leq \|D_{y+h}^2\phi - D_y^2\phi\|\|h\|^2$ et que ϕ est de classe C^2, (5) devient

$$D_{y+h}\phi(h) - D_y\phi(h) = D^2\phi_y(h,h) + o(\|h\|^2)$$

$J(y) = \mathbb{E}[d_\phi(X,y)] = \int_E d_\phi(x,y)\mu(dx) = \int_E(\phi(x) - \phi(y) - D_y\phi(x-y))\mu(dx)$

$$\begin{aligned} &J(y+h) - J(y) \\ &= \int_E(-\phi(y+h) - D_{y+h}\phi(x-y-h))\mu(dx) + \int_E(\phi(y) + D_y\phi(x-y))\mu(dx) \\ &= -\int_E(\phi(y+h) - \phi(y))\mu(dx) - \int_E(D_{y+h}\phi(x-y-h) - D_y\phi(x-y))\mu(dx) \end{aligned}$$

Or, on a $\phi(y+h)-\phi(y) = D_y\phi(h)+\varepsilon(y,h)\|h\|$ et

$$\begin{aligned}
&D_{y+h}\phi(x-y-h) - D_y\phi(x-y)\\
&\quad= D_{y+h}\phi(x-y-h) - D_y\phi(x-y-h) + D_y\phi(x-y-h) - D_y\phi(x-y)\\
&\quad= D_y^2\phi(x-y-h,h) + \varepsilon(y,h)\|h\| + D_y\phi(x-y) - D_y\phi(h) - D_y\phi(x-y)\\
&\quad= D_y^2\phi(x-y,h) - D_y^2\phi(h,h) - D_y\phi(h) + \varepsilon(y,h)\|h\|
\end{aligned}$$

Donc

$$\begin{aligned}
&J(y+h) - J(y)\\
&\quad= -\int_E (D_y\phi(h)+\varepsilon(y,h)\|h\|)\mu(dx)\\
&\qquad -\int_E (D_y^2\phi(x-y,h) - D^2\phi_y(h,h) - D\phi_y(h) + \varepsilon(y,h)\|h\|)\mu(dx)\\
&\quad= -\int_E (D_y^2\phi(x-y,h) - D_y^2\phi(h,h) + \varepsilon(y,h)\|h\|)\mu(dx)
\end{aligned}$$

Il existe $m \geq 0$ tel que $\|D_y^2\phi(h,h)\| \leq m\|h\|^2$, donc on a

$$\lim_{\|h\|\to 0} \frac{\|J(y+h) - J(y) + \int_E D_y^2\phi(x-y,h)\mu(dx)\|}{\|h\|} = 0$$

D'où $D_yJ(h) = -\int_E D_y^2\phi(x-y,h)\mu(dx) = -D_y^2\phi(\int_E (x-y)\mu(dx), h)$. Or, la condition $D_yJ(h) = 0$ pour tout $h \in E$ est une condition nécessaire pour que J ait un extremum en y. On a $D_yJ(h) = 0 \Leftrightarrow D_y^2\phi(\int_E (x-y)\mu(dx), h) = 0$, et en particulier,

$$\begin{aligned}
D_yJ\left(\int_E (x-y)\mu(dx)\right) = 0 \quad &\Leftrightarrow \quad D_y^2\phi\left(\int_E (x-y)\mu(dx), \int_E (x-y)\mu(dx)\right) = 0\\
&\Leftrightarrow \quad \int_E (x-y)\mu(dx) = 0\\
&\qquad\quad \text{car } D_y^2\phi(\cdot,\cdot) \text{ est coercive}\\
&\Leftrightarrow \quad y = \mathbb{E}[X]
\end{aligned}$$

Si J admet un extremum, c'est en $y = \mathbb{E}[X]$. On va voir qu'il s'agit d'un minimum. Pour cela, montrons que $D_y^2J(\cdot,\cdot)$ est coercive.

$$\begin{aligned}
&D_{y+u}J(h) - D_yJ(h)\\
&\quad= -\int_E (D_{y+u}^2\phi(x-y-u,h) - D_y^2\phi(x-y-u,h)\\
&\qquad\qquad + D_y^2\phi(x-y-u,h) - D_y^2\phi(x-y,h))\mu(dx)\\
&\quad= -\int_E D_y^3\phi(x-y-u,h,u) + \varepsilon(y,h,u)\|u\| - D_y^2\phi(u,h)\mu(dx)\\
&\quad= -\int_E D_y^3\phi(x-y,h,u) - D_y^3\phi(u,h,u) + \varepsilon(y,h,u)\|u\| - D_y^2\phi(u,h)\mu(dx)
\end{aligned}$$

On a

$$\lim_{\|u\|\to 0} \frac{\|D_{y+u}J(h) - D_yJ(h) + \int_E D_y^3\phi(x-y,h,u)\mu(dx) - \int_E D_y^2\phi(h,u)\mu(dx)\|}{\|u\|} = 0$$

Donc

$$D_y^2J(h,u) = -\int_E D_y^3\phi(x-y,h,u)\mu(dx) + \int_E D_y^2\phi(h,u)\mu(dx)$$

Prenons $u = h$, et y réalisant l'extremum. On a alors

$$\begin{aligned} D_y^2J(h,h) &= -\int_E D_y^3\phi(x-y,h,h)\mu(dx) + \int_E D_y^2\phi(h,h)\mu(dx) \\ &= -D_y^3\phi\left(\int_E (x-y)\mu(dx), h, h\right) + \int_E D_y^2\phi(h,h)\mu(dx) \\ &= \int_E D^2\phi_y(h,h)\mu(dx) \\ &\geq \int_E c\|h\|^2\mu(dx) \\ &= c\|h\|^2 \end{aligned}$$

Ainsi, $D_y^2J(\cdot,\cdot)$ est coercive, et $y = \mathbb{E}[X]$ est un minimum. □

Lemme 5.3 (Meilleur ensemble de centres). *Soit q un quantificateur de partition associée $\{S_i\}_{i=1}^k$ avec $\mu(S_i) > 0$ pour $i = 1\ldots,k$. Si q' est un quantificateur de même partition dont les centres sont $y_1', \ldots, y_k'$ sont définis par*

$$y_i' \in \arg\min_{y\in E} \mathbb{E}[d_\phi(X,y)|X \in S_i], i = 1,\ldots,k$$

alors

$$W(\mu,q') \leq W(\mu,q)$$

Démonstration. On a

$$\begin{aligned} W(\mu,q) &= \mathbb{E}d_\phi(X,q(X)) \\ &= \sum_{i=1}^k \mathbb{E}[d_\phi(X,y_i)|X\in S_i]\mu(S_i) \\ &\geq \sum_{i=1}^k \mathbb{E}[d_\phi(X,y_i')|X\in S_i]\mu(S_i) \\ &= \mathbb{E}d_\phi(X,q'(X)) \\ &= W(\mu,q') \end{aligned}$$

□

6 Bibliographie

[1] Banerjee, A., Merugu, S., Dhillon, I.S., et Ghosh, J. (2005) Clustering with Bregman Divergences, *Journal of Machine Learning Research*, vol. 6, 1705-1749.

[2] Biau, G., Devroye, L. and Lugosi, G. (2005). On the performance of clustering in Hilbert spaces, *IEEE Transactions on Information Theory*.

[3] Billingsley, P. (1979). *Probability and Measure*, Wiley Series in Probability and Mathematical Statistics, J. Wiley and Sons.

[4] Boucheron, S., Bousquet, O. et Lugosi, G. (2005). Theory of classification : A survey of some recent advances. *ESAIM : Probability and Statistics*, Vol. 9, 323-375.

[5] Bregman, L.M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, *USSR Computational Mathematics and Mathematical Physics*, vol. 7, 200-217.

[6] Frigyik, B.A., Srivastava, S. et Gupta, M.R. (2006). Functional Bregman Divergence and Bayesian Estimation of Distributions.

[7] Gelfand, I.M. et Fomin, S.V. (1963). *Calculus of variations*, Prentice-Hall, Englewood Cliffs, New Jersey.

[8] Giesy, D.P. (1976). Strong laws of large numbers for independent sequences of Banach space-valued random variables, *Probability in Banach Spaces, Oberwolfach 1975*, Lectures Notes in Mathematics, Springer-Verlag.

[9] Ledoux, M. et Talagrand, M. (1991). *Probability in Banach Spaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer.

[10] Linder, T. (2001). Learning-Theoretic Methods in Vector Quantization, Lecture notes for the Advanced School on the Principles of Nonparametric Learning, Udine, Italie, 9-13 juillet 2001.

[11] Massart, P. (2007). *Concentration Inequalities and Model Selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003, Lecture Notes in Mathematics, Springer.

[12] McDiarmid, C. (1989). On the method of bounded differences, in *Surveys in Combinatorics 1989*, 148-188, Cambridge University press, Cambridge.

[13] Pollard, D.B. (2002). *A User's Guide to Measure Theoretic Probability*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

[14] Ramsay, J.O. et Silverman, B.W. (1997). *Functional Data Analysis*, Springer Series in Statistics, Springer.

[15] Rachev, S.T. et Rüschendorf, L. (1998). *Mass Transportation Problems*, vol. I, Probability and its Applications, Springer.