Projets informatiques
L3 EURIA

10 octobre 2024

Consignes :

— Les soutenances dureront 20 minutes (10 minutes de présentation et 10 minutes de questions). Elles auront lieu le **9 décembre**.

— Envoyer un premier compte-rendu avant le **9 novembre**. Ce comptes-rendu décrira les travaux effectués ainsi qu'un planning prévisionnel pour la suite du travail.

— N'attendez pas le dernier moment pour travailler sur les projets et n'hésitez pas à poser des questions.

— **N'oubliez pas de respecter la charte anti-plagiat de l'UBO** disponible via le lien ci-dessous
`http://ubodoc.univ-brest.fr/wp-content/uploads/2014/12/charte-antiplagiat-ubo.pdf`

— Le rapport (environ 4 pages au format .pdf) et les programmes commentés devront être déposés au plus tard le **3 décembre** sur la plateforme Moodle de l'EURIA :
`https://moodlesciences.univ-brest.fr/moodle/course/view.php?id=1083`

— l'équipe pédagogique ayant bien conscience de la possibilité de recourrir à des générateurs de code automatique (via des agents conversationnels), il est impératif que tout membre du groupe s'approprie le code rendu. En particulier, lors de la soutenance tout membre devra être capable d'expliquer précisément toute partie du code rendu questionné par le jury.

1. Classification et représentation de données par cartes de Kohonen :

— *Langage :* R, R-Shiny

— *Sujets abordés :* analyse de données, algorithme stochastique

1.a) Introduction

L'analyse de données multidimensionnelles porte typiquement sur l'étude de n individus déterminés par p variables (quantitatives ou qualitatives) : un individu est un élément d'un espace de dimension p.

Les méthodes factorielles (A.C.P. et dérivées), qui sont en fait des méthodes de projection de l'Algèbre linéaire permettent des *représentations* graphiques des données en dimension 1,2 ou 3. Il est cependant difficile de construire des classes de proximité à partir des projections si les individus ne sont pas représentables sans perte d'information dans un espace de dimension

≤ 3. Deux individus dont les projections sont proches ne sont pas toujours proches dans l'espace initial de dimension p.

D'un autre côté, nous pouvons distinguer deux familles de méthodes de *classification* :

- Classification hiérarchique : processus itératif d'agrégation des classes les plus voisines. Initialement, chaque individu est une classe. On considère les n classes et on regroupe les 2 plus proches (pour une distance choisie) pour former une nouvelle classe (de 2 individus). Et on itère sur les $n-1$ classes obtenues...jusqu'à l'obtention d'une seule classe. Ceci donne lieu à une représentation par un arbre. On peut ensuite choisir le nombre de classes qui paraît le mieux adapté en sélectionnant un niveau de regroupement.

- Classification non hiérarchique : le nombre de classes est fixé à priori et on attribue sa classe à chaque individu par un algorithme convergeant vers une répartition minimisant l'inertie intra-classe (i.e. les sommes des distances de chaque individu d'une classe au centre de la classe). Exemple : algorithme de moyennes mobiles, ou k-means.

Pour ces méthodes de classification, deux points d'une même classe sont proches dans l'espace initial, mais comment représenter globalement les classes en conservant la topologie initiale des données ? Il n'y a pas de notion de voisinage des classes.

L'algorithme de Kohonen tente de jumeler *Représentation* et *Classification*. Teuvo Kohonen a proposé dès 1982 un algorithme dont la fonction principale est de faire correspondre les éléments de l'espace des entrées avec des unités ordonnées sur une carte – une représentation graphique (de dimension 1, 2 ou 3) où chaque unité est entourée de ses voisines (pour une distance prédéfinie).

Le résultat est une fonction de l'espace des entrées vers l'ensemble des unités, telle que les images de deux éléments voisins au sens d'une certaine distance dans l'espace des entrées sont la même unité ou des unités voisines sur la carte.

A noter que T.Kohonen a proposé cet algorithme dans le cadre de ses travaux sur la modélisation mathématique du fonctionnement des neurones biologiques. L'algorithme a été conçu en 1982 comme une modélisation de la formation automatique de cartes dans les zones sensorielles du cortex même si depuis, l'étude mathématique de l'algorithme et ses applications – notamment en Analyse des données – l'ont éloigné de son cadre biologique originel. Par exemple, considérons les connexions nerveuses des cellules rétiniennes vers le cortex cérébral traitant les informations visuelles (le tectum). Si on imagine la rétine comme une grille de dimension 2, un fait important est que la topologie de la rétine est préservée par l'ensemble des connexions dans le sens que deux cellules proches sur la rétine sont connectées à deux cellules proches dans le cortex. Il n'y a aucune raison que ceci se réalise spontanément à la naissance, et on peut penser que ce phénomène est dû à un processus d'auto-organisation et de sélection gouverné par les activités spontanées des neurones cérébraux.

C'est précisément cette idée que T.Kohonen a voulu adapté lorsqu'il a défini son algorithme d'auto-organisation, ajoutant un concept important aux modèles connexionnistes antérieurs qu'il connaissait très bien (modèles de mémoire associative de D.Gabor, J.Hopfield, etc...). D'un point de vue algorithmique, le processus d'auto-organisation se déroule par une mise à jour locale des connections selon des règles complémentaires de compétition et de coopération, à chaque présentation d'un prototype.

1.b) Notations et définitions du modèle

L'**espace des données** est un sous-ensemble borné convexe $X \subset \mathbb{R}^p$. Nous considérons que

$\mathbb{R}^p$ est muni d'une norme $\|.\|$ (par exemple la norme associée à la distance Euclidienne).

$\Rightarrow$ On considére un échantillon $(x(1), \ldots, x(t), \ldots)$, issu d'observations successives dans X (approche statistique), ou une suite de réalisations de variables aléatoires indépendantes de même loi de probabilité μ à valeurs dans X (approche probabiliste).

Le **réseau** est formé de n unités (ou neurones) disposés selon une topologie déterminée :

- à $d = 1, 2$ ou 3 dimensions : une ligne, un carré ou un cube,

- selon un maillage dont la structure de voisinage est déterminée par une fonction de voisinage.

On représente les unités par un sous-ensemble I de $\mathbb{Z}^d$, et la fonction de voisinage est une fonction Λ définie sur $I \times I$,

- symétrique (i.e. $\Lambda(i, j) = \Lambda(j, i)$),

- dépendant seulement d'une distance D sur I, de norme associée $|\ .\ |$.

- décroissant avec la distance : $\Lambda(i, j)$ tend vers 0 quand $D(i, j) = |\ i - j\ | \longrightarrow +\infty$.

On convient souvent que $\Lambda(i, i) = 1$. Quand $\Lambda(i, j) = 1$, on dit que i et j sont fortement connectés, et quand $\Lambda(i, j) = 0$, i et j sont totalement disconnectés et n'ont pas d'interaction.

Chaque unité i est dotée d'un vecteur d'état $W_i(t) \in \mathbb{R}^p$ pointant dans l'espace des donnéees, et susceptible d'être modifié.

L'état du réseau à l'instant t est donné par $(W_i(t), i \in I)$.

$\Rightarrow$ L'objectif est de trouver des vecteurs W_i ayant des propriétés de :

- **Quantification** : le nombre de vecteurs W_i dans une région A de X est approximativement proportionnel à $\mu(A)$,

- **Organisation** : deux unités i et j proches (i.e. $\Lambda(i, j) \sim 1$) ont des vecteurs W_i et W_j proches.

Une fois de tels vecteurs W_i déterminés, ils sont utilisés pour définir la classe d'un élément $x \in X$ quelconque : on attribue à x la classe i^* telle que

$$i^*(x, W) = \text{argmin}\{\|x - W_i\|, i \in I\}.$$

Etudions maintenant l'algorithme de Kohonen qui détermine des vecteurs W_i par une méthode d'apprentissage.

1.c) L'Algorithme de Kohonen

- Les vecteurs $W_i(t = 0)$ sont initialisés aléatoirement.

- A l'instant t, l'état du réseau est donné par $W(t) = (W_i(t), i \in I)$. Un vecteur de l'espace des données $x(t + 1)$ est choisi aléatoirement.

- La **phase de compétition** désigne l'unité gagnante :

$$i^*(x(t+1), W(t)) = \text{argmin}\{\|x(t+1) - W_i(t)\|, i \in I\} \qquad (\mathbf{1})$$

(Dans le cas où plusieurs unités minimisent cette distance, on convient d'une règle, par exemple le premier indice i pour l'ordre lexicographique sur $\mathbb{Z}^d$).

• La **phase de coopération** modifie les vecteurs W_i :

$$\forall j \in I, W_j(t+1) = W_j(t) - \varepsilon_t \, \Lambda_t(i^*, j) \, (W_j(t) - x(t+1)) \qquad (\mathbf{2})$$

On poursuit l'algorithme tant que t est inférieur à une valeur seuil M fixée à priori, ou on peut imposer une condition d'arrêt alternative ou supplémentaire s'il n'y a plus d'amélioration notable. Seules les unités proches du neurone gagnant ont leur vecteur W_i modifié : notion de coopération.

Les paramètres essentiels sont :

- La dimension p de l'espace des données.

- La topologie du réseau.

- La fonction de voisinage Λ_t, constante ou dépendant du temps.

Exemples : $\Lambda_t(i,j) = \mathbf{1}_{\{|i-j| \leq k\}}, (k = 1, \text{ ou } 2)$.

$\Lambda_t(i,j) = g(\mid i - j \mid)$, où g est une fonction en cloche.

$\Lambda_t(i,j) = g(\frac{\mid i - j \mid}{\lambda(t)})$, où $\lambda(t) \to 0$, quand $t \to +\infty$.

Un exemple courant est : $\Lambda_t(i,j) = \exp(-\frac{\mid i - j \mid^2}{2\sigma(t)^2})$, où $\sigma(t) = \sigma_i \left(\frac{\sigma_f}{\sigma_i}\right)^{t/M}$, avec $\sigma_i > \sigma_f > 0$ (par exemple $\sigma_i = 5$ et $\sigma_f = 0,2$).

- Le paramètre d'adaptation ε_t, à valeurs dans $]0,1[$, qui peut être constant ou décroissant en t. Par exemple, choisir ε_t tels que $\sum_{t \geq 0} \varepsilon_t = +\infty, \sum_{t \geq 0} \varepsilon_t^2 < +\infty$, comme $\varepsilon_t \sim 1/t$ (conditions de Robbins-Monro issues des Algorithmes stochastiques).

Un autre exemple courant est : $\epsilon(t) = \epsilon_i \left(\frac{\epsilon_f}{\epsilon_i}\right)^{t/M}$, avec $\epsilon_i > \epsilon_f > 0$ (par exemple $\epsilon_i = 0,1$ et $\epsilon_f = 0,005$).

- La loi de probabilité μ.

1.d) Applications

Cet algorithme et ses très nombreuses variantes connaissent une multitude d'applications, par exemple : analyse des données (classification), traitement du signal, reconnaissance d'images et acoustiques (paroles,...), prévision de séries chronologiques, robotique, optimisation (le problème célèbre du voyageur de commerce).

On pourra consulter notamment l'article de M. Cottrell *et al* (2003) où sont présentés quelques exemples : analyse et comparaison de pays repérés par des variables socio-économiques, courbes de consommation électrique, démographie dans des communes de la vallée du Rhône, profils de consommateurs au Canada, segmentation du marché du travail en France, etc. L'avantage de cette méthode de classification est qu'elle permet de réprésenter graphiquement les observations en respectant la topologie du réseau. On associe ainsi à chaque unité les observations appartenant à cette classe. Lorsque les données et les classes sont très nombreuses, il est possible de réaliser une deuxième classification sur les vecteurs représentant les classes, afin d'obtenir un niveau de classification plus grossier, qui sera plus facile à interpréter. De nombreuses variantes de cet algorithme ont été proposés. En particulier, il est possible de l'appliquer dans le cas de variables qualitatives, en l'associant à une analyse factorielle des correspondances (Cottrell *et al*, 2003]).

Dans le cadre de ce projet, il est demandé d'implémenter cet algorithme et de réaliser une interface R-Shiny pour rendre l'application interactive, avec des animations proposées suite à des choix de paramètres. Pour l'application R-Shiny, on considérera des données en dimension 2, et une grille en dimension 1 ou 2. L'avantage est qu'on pourra superposer graphiquement l'espace des données et la grille et représenter les vecteurs W_i dans l'espace des données. L'objet de l'interface est d'être un outil pédagogique permettant de visualiser facilement le mécanisme d'auto-organisation des vecteurs W_i.

1.e) Bibliographie

M.Cottrell, S. Ibbou, P. Letremy,, P. Rousset, "*Cartes organisées pour l'analyse exploratoire des données et la visualisation*", Journal de la Société Française de Statistique, n° 144, p. 67-106, (2003). (`https://arxiv.org/abs/math/0611422`)