Proposition de stage de Master 2 :
Partitionnement de données et quantification sur des varifolds

Claire Brécheteau, Stéphane Guillermou

7 février 2024

Description du projet La classification de données est une des branches majeures de la statistique. La classification consiste à séparer des données en groupes disjoints de sorte que des données au sein d'un même groupe soient le plus similaires possibles et que des données de groupes distincts soient assez différentes. La classification permet d'affecter une unique étiquette à chaque donnée, indiquant son groupe. Dans le cadre de la classification non supervisée, seules les données sont accessibles, au sens où aucune étiquette n'est connue à l'avance. C'est dans ce cadre que ce stage se situe. Les données considérées sont des données dans $\mathbb{R}^d$. Elles sont éventuellement issues d'une mesure supportée sur une variété immergée dans $\mathbb{R}^d$ avec potentiellement des points doubles.

Une méthode classique en statistiques pour partitionner des données est le critère des k-means. Lorsqu'il est associé à une mesure de probabilité sur $\mathbb{R}^d$, il est défini pour un dictionnaire de k centres $\mathbf{c} = (c_1, c_2, \dots, c_k)$ par

$$R(\mathbf{c}) = \int \min_{i=1..k} \|x - c_i\|^2 \, dP(x).$$

Les centres $\mathbf{c}^*$ minimisant ce critère sont des représentants de la mesure P. Ces centres optimaux induisent une décomposition de l'espace $\mathbb{R}^d$ en cellules de Voronoï.

Lorsque la mesure P n'est pas connue, mais que l'on a accès à un n-échantillon d'observations indépendantes de cette loi P, $X_1, X_2, \dots, X_n$, alors les centres empiriques optimaux $\hat{\mathbf{c}}$ sont ceux qui minimisent le critère empirique,

$$R_n(\mathbf{c}) = \frac{1}{n} \sum_{j=1}^{n} \min_{i=1..k} \|X_j - c_i\|^2.$$

Ils induisent également une décomposition de l'espace en cellules, et donc un partitionnement des données en k groupes.

Des versions robustes à certains types de bruit dans les données ont été étudiées dans ce contexte [3], [1].

Lorsque le support de la mesure P est une variété immergée qui possède des points doubles (par exemple, le symbole infini), une application des k-means pourrait générer des groupes contenant des points issus de deux parties potentiellement très éloignées dans la variété originelle. Dans ce travail, il s'agit de tenir compte de l'espace tangent associé aux données pour se prémunir d'un tel problème.

Un bon concept pour pallier cette difficulté est celui de varifold, [2, 4].

L'étudiant·e sera amené·e, après lecture des différentes ressources bibliographiques, à effectuer des avancées mathématiques dans la justification théorique statistique de l'algorithme des k-means sur une varifold, et/ou de ses versions robustes à différentes formes de bruit. Il/elle sera amené·e à coder en Julia ou tout autre langage une telle méthode et à l'appliquer sur des données simulées. Ce code pourra être intégré par exemple à la bilbiothèque en construction `GeometricClusterAnalysis.jl` en Julia.

Profil du/de la candidat·e Nous recherchons un ou une étudiant·e de niveau M2, issu·e d'une formation de statistiques, probabilités ou mathématiques fondamentales.

Détails

- Encadrants : Claire Brécheteau, Stéphane Guillermou
- Début : à partir du 1er mars 2024
- Durée : 4-6 mois
- Localisation : Laboratoire de Mathématiques Jean Leray, Nantes Université

Comment Candidater Les candidat·e·s devront envoyer un mail avec CV et lettre de motivation à Claire Brécheteau (claire.brecheteau@ec-nantes.fr) et Stéphane Guillermou (stephane.guillermou@univ-nantes.fr)

Références

[1] C. Brécheteau and C. Levrard. A k-points-based distance for robust geometric inference. *Bernoulli*, 26(4) :3017–3050, 2020.

[2] B. Buet, G. P. Leonardi, and S. Masnou. A varifold approach to surface approximation. *Arch. Ration. Mech. Anal.*, 226(2) :639–694, 2017.

[3] J. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means : An attempt to robustify quantizers. *The Annals of Statistics*, 25 :553–576, 04 1997.

[4] R. Tinarrage. Recovering the homology of immersed manifolds. *Discrete Comput. Geom.*, 69(3) :659–744, 2023.