Université de Paris

Data Science avancée

Apprentissage non supervisée

1. Jeux de données

Trois jeux de données (Classic 3, Classic4 et BBC) labellisés sont fournis et serviront d'évaluation de différentes approches.

- Classic 3 : Il contient 3893 documents catégorisés en 3 classes.
- Classic4: Il contient 7095 documents catégorisés en 4 classes.
- BBC: Il contient 2225 articles de presse catégorisés en 5 classes.

2. Travaux à réaliser

Les jeux de données proposés sont utilisés pour étudier les représentations textuelles vues en cours, à savoir Word2vec, GloVe[1]. Ces données serviront de domaines d'application de méthodes vues en cours ou à découvrirr. Le projet comportera deux parties différentes : la première concerne principalement la réduction de la dimension à laquelle est ajoutée ensuite une tâche de clustering (approche tandem), la seconde se focalise sur l'obtention du clustering via une approche simultanée combinant les deux tâches simultanément.

3. Partie 1 : Approche Tandem

Une fois les représentations obtenues, il vous sera demandé de réaliser une étude comparative de différentes méthodes (vues ou non en cours) de réduction de dimension (PCA, t-SNE, UMAP, Autoencodeurs) et de clustering (Kmeans++, Kmedoids, spherical Kmeans, CAH avec différents critères d'agrégation) dans l'espace réduit et l'espace d'origine.

1. A l'aide des métriques accuracy, NMI et ARI, évaluer le clustering à partir de l'espace d'origine et l'espace réduit sur labse du vrai nombre de classes.
2. Une interprétation des classes doit être réalisée.
3. Une étude sur l'estimation du nombre de classes erst à réaliser à partir de critères disponibles par exemple dans le package **NbClust.** https://www.jstatsoft.org/article/view/v061i06

Dans cette partie, on doit disposer de tableaux synthétiques, de visualisations en 2d ou 3d et des commentaires pertinents de chaque table et figure. A noter que le code de ces méthodes est disponible enR et Python

4. Partie 2 : Approche jointe/simultanée

Dans cette partie et contrairement à la partie 1 (approche Tandem), il s'agit d'appliquer et d'évaluer des méthodes combinant simultanément les méthodes de la réduction de dimension et le clustering.

1. Reduced k-means et Factorial k-means [3, 4] https://cran.r-project.org/web/packages/clustrd/clustrd.pdf
2. Deep k-means (DKM) [1] https://github.com/MaziarMF/deep-k-means

Comme dans Partie 1, on doit disposer de tableaux synthétiques, de visualisations en 2d ou 3d et des commentaires pertinents de chaque table et figure. En plus, des commentaires comparatifs de Partie 1 et Partie 2 seront nécessaires. A noter que le code de ces méthodes est disponible.

5. Rendus du projet en deux étapes

- @AMSD Le retour du projet est programmé pour 27 mars à minuit.
- @MLSD Le retour du projet est programmé pour 10 avril à minuit.

6. Envois des projets

Les envois sont à adresser mohamed.nadif@u-paris.fr (en spécifiant dans le sujet de votre message Data Science avancée)

References

[1] M. M. Fard, T. Thonet, and E. Gaussier. Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognition Letters, 138:185–192, 2020.

[2] Stephen L France and Ulas Akkucuk. A review, framework, and r toolkit for exploring, evaluating, and comparing visualization methods. Vis. Comput., 37(3):457–475, 2021.

[3] M. Vichi and H. Kiers. Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis, 37(1):49–64, 2001.

[4] M. Yamamoto and H. Hwang. A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika, 41(1):115–129, 2014.

Note Importante : Ce projet est à réaliser par binôme ou seul(e). Le rendu doit être sous forme notebook comportant le code utilisé tout en commentant les arguments de vos fonctions et les résultats. Toute ressemblance entre deux rendus sera sanctionnée. Ce projet servira d'une bonne préparation de l'examen sous forme de QCM qui aura lieu après remise des projets. Cette date vous sera communiquée dans les prochains jours.

[1]Utiliser la version https://nlp.stanford.edu/data/glove.840B.300d.zip de GloVe en convertissant le modèle GloVe en un format word2vec avec la fonction : https://radimrehurek.com/gensim/scripts/glove2word2vec.html