Université Paris Cité

Data Science avancée

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 https://drive.google.com/file/d/1SpcExC4Fs2H-_9hdem4uz_6bd8L12RgF/view?usp=drivesdk
- Classic 4 https://drive.google.com/file/d/1fMeNQm4fF1NpgTe2WVO3kQP16p_N6fYw/view?usp=drivesdk
- BBC https://drive.google.com/file/d/1vK6cG924aVxeVL7di_xfGHibL7TwNsjg/view?usp=drivesdk

2. Travaux à réaliser

Les jeux de données proposés sont utilisés pour étudier les représentations textuelles Word2vec, GloVe[1]. Ces données serviront de domaines d'application de méthodes vues en cours ou à découvrir. 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 la base du vrai nombre de classes. Ces métriques dites « externes » serviront juste à comparer les performances des différents algorithmes sur des données labélisées connues.
2. Une interprétation des classes doit être réalisée.
3. Une étude sur l'estimation du nombre de classes est 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 en R 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 du clustering.

1. Reduced k-means [2] 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

Note Importante : Ce projet est à réaliser par binôme ou seul(e).

- *@AMSD Le retour du projet est programmé pour 5 mai à minuit si le binôme est constitué exclusivement de AMSD.*
- *@MLSD Le retour du projet est programmé pour 12 mai à minuit si le binôme est constitué exclusivement de MLSD.*
- *Dans le cas contraire le retour est programmé pour le 8 mai à minuit.*

6. Envois des projets

Les projets sont à déposer sur Slack à mon adresse.

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

7. Consignes à respecter rigoureusement

1. ***Un notebook comportant le code utilisé tout en commentant les arguments de vos fonctions et les résultats.***
2. ***Une vidéo de 10mn présentant votre travail***
3. ***Un poster***

Toute ressemblance entre deux rendus sera sanctionnée. **Aucun retard ne sera toléré.**

1Utiliser 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