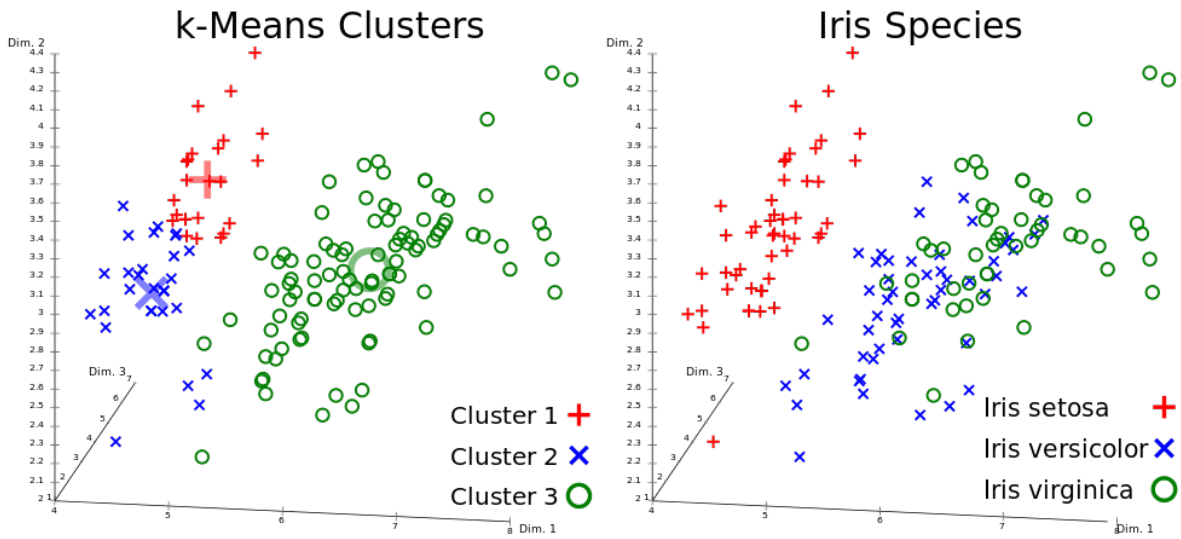
Big Data. TP Projet. Clustering

Natalia Kharchenko, Sergey Kirgizov

Polytech Dijon



L'objectif est de se familiariser avec les techniques du clustering et les techniques de visualisation. Il y a deux projets : implémentation de k-Means et exploration des jeux de données. Vous pouvez choisir de réaliser soit un de ces projets, soit les deux projets.
Vous pouvez réaliser les projets soit seul(e) soit en binôme.
Vous allez présenter les résultat lors des prochaines séances de TP.

Technologies recommandées : `python`, `numpy`, `pandas`, `sklearn`, `matplotlib`. N'hésitez pas à utiliser d'autres langages de programmation ou technologies de visualisation si vous en avez envie.

1 Le projet de codage : implémentation de k-Means

L'objectif est d'implémenter l'algorithme de clustering k-Means afin d'acquérir l'intuition de son fonctionnement. N'hésitez pas à réutiliser votre travail du TD 3 (`https://kirgizov.link/teaching/polytech-dijon/bigdata/TD-3-kMeans.pdf`) dans ce projet.

1. Créez et visualisez un ensemble de données artificielles pour tester votre algorithme. Il faut varier la dimensionnalité et la distribution des points. Toutes les implémentations du ce projet il faut tester au moins avec les données de 1D, 2D, 3D et 4D et 2 distributions différentes (gaussianne + une autre).
2. Implémenter l'algorithme classique de k-means. Testez sur les données de l'exercice précédent et visualisez les clusters. Vérifiez l'évolution de la somme des carrés des erreurs (Sum of Squared Errors – SSE) avec les itérations de l'algorithme.
3. Implémenter k-means++. Comparez la vitesse de convergence et la qualité de la solution finale avec l'algorithme classique (l'algorithme de Lloyd).
4. Implémenter mini-batch k-means. Testez sur un échantillon plus large (10000+ points).
5. Testez différentes valeurs du paramètre k correspondant au nombre de clusters. Choisissez le meilleur k en utilisant la méthode du coude ou de la silhouette.

ASTUCE : Si vous rencontrez des problèmes de traitement ou de visualisation de données multidimensionnelles, vous pouvez utiliser les techniques de la réduction dimensionnelle, par exemple PCA.

ASTUCE : Si vous rencontrez des problèmes de visualisation de grand nombre des points, utilisez des techniques de binning rectangulaire ou hexagonale.

2 Exploration des jeux de données avec clustering

L'objectif est d'explorer un ensemble de données du monde réel en utilisant des algorithmes du clustering et des techniques de réduction de la dimension.

0. Sélectionnez un jeu de données pour l'analyse. Vous pouvez choisir le jeu de données à partir de l'une des options mentionnées à la fin de ce TP. La note maximale du projet dépend du choix des données. Pour les toy datasets de `sklearn` la note maximale est 16/20. Pour les autres jeux de données la note maximale est 20/20.
1. Effectuez le traitement préliminaire des données (sélection des propriétés intéressantes, conversion des données pour le calcul de distances, etc).
2. Effectuez une analyse des données en utilisant l'algorithme k-means. Testez différentes valeurs du paramètre k correspondant au nombre de clusters. Choisissez le meilleur k en utilisant la méthode du coude ou de la silhouette.
3. Présenter graphiquement les résultats.
4. Pourriez-vous interpréter (décrire chaque groupe en un mot ou une phrase) les clusters trouvés par l'algorithme k-means ? Essayez de comparer les résultats du clustering k-means avec de vraies classes d'objets, s'ils existent dans votre jeu de données.

Jeux de données

1. Toy datasets de `sklearn` `https://scikit-learn.org/stable/datasets/toy_dataset.html`
2. Qualité du vin, 4 898 instances `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`
3. Génomique, fréquences d'utilisation des codons dans ADN, 13 028 instances `http://archive.ics.uci.edu/ml/datasets/Codon+usage`
4. Chiffres manuscrits. Données optiques. 5620 instances `http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits`
5. Chiffres manuscrits. Tablette sensible à la pression. 10 992 instances `http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits`
6. HIGGS, 11 000 000 instances `http://archive.ics.uci.edu/ml/datasets/HIGGS`
7. Scrutins de l'assemblée nationale `http://data.assemblee-nationale.fr/travaux-parlementaires/votes` Exemple de clustering k-means : `https://www.data.gouv.fr/fr/reuses/clustering-k-means-des-deputes-par-leurs-votes/`

Vous pouvez utiliser également tout autre jeu de données légalement disponible sur le réseau Internet, par exemple les jeux de données provenant de

- UC Irvine Machine Learning Repository : `http://archive.ics.uci.edu/ml/datasets.php?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table`
- Plateforme ouverte des données publiques françaises : `https://www.data.gouv.fr/fr/`
- Plateforme web organisant des compétitions en science des données Kaggle `https://www.kaggle.com/datasets`