University of Vienna
Faculty of Computer Science
Prof. Claudia Plant
Martin Perdacher

Scientific Data Management

SS 2019

General Remarks:

- This is one of three programming assignments in this lecture. For each assignment you could earn 100 points.
- The deadline is 4th of April 2019. No deadline extensions are given.
- If you have problems do not hesitate to contact the tutor or post a question on the Moodle system.
- Pack up your java code, your results, a jar-file and the documentation in a .zip file with the following name: group(group number).zip file (e.g. group01.zip) and upload it to the Moodle system.
- Only one team-member submits the zip file in the moodle system.
- Do not use a built-in implementation of K-means, but T-SNE, PCA or any sampling library is fine.

External resources:

- NMI-implementation in Java:
`https://gist.github.com/perdacherMartin/76689fdf2c950fbeba6b013d09906de4`
- Leaderboard: `https://app.webjets.io/lib/1551876706478-193d`
- Skin segmentation dataset `https://archive.ics.uci.edu/ml/datasets/skin+segmentation`
- HTRU2 dataset `https://archive.ics.uci.edu/ml/datasets/HTRU2`.

Task 1-1 K-means

Implement the K-means algorithm in Java. Form groups of three or four persons. Your solution should contain the following functionalities:

- (30 points) Implement two different update strategies, what have been introduced in the lecture, where the algorithm updates
  (a) each round (Lloyd) or
  (b) immediately after assignment of each point (Mac Queen)
- (30 points) Strategies for the initialisation. Besides random initialization techniques, implement at least two other techniques. There are very different ideas for careful seeding[1] or using sampling approaches[2]. Most of them are summarized in a survey on initializaiton methods [3]. Measure the time and the quality (NMI) of your experiments (average of 100 runs) and report them on our leaderboard for the two datasets (HTRU2 and skin-segmentation).
- (20 points) Discussion: Try to argue and to discuss the convergence and the quality of different algorithms or initialization techniques you have implemnted. Try to support your claims with visualisations. (this could be anything, NMI-boxplot, scatter plot with dimensional reduction techniques (T-SNE or PCA)).
- (20 points) Documentation: Write a documentation, which describes your implementation. Be aware, that your results needs to be reproduceable. Comment also on software prerequisites.

Literatur

[1] Arthur, D., and Vassilvitskii, S. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007* (2007), pp. 1027–1035.

[2] Bachem, O., Lucic, M., Hassani, S. H., and Krause, A. Fast and provably good seedings for k-means. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain* (2016), pp. 55–63.

[3] Celebi, M. E., Kingravi, H. A., and Vela, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *CoRR abs/1209.1960* (2012).