**PROJECT – Evaluation of Clustering Algorithms**

Sergei Rogov

U231N0051

University of Nicosia

COMP-344 (Section 01): Machine Learning and Data Mining II

## Instructor: Dr Ioannis Katakis

January 5, 2024

Goals:

Summary:

**Datasets**

**1-1.5**

**Dataset 1**

[**https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data**](https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data)

**Dataset 2**

[**https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset/data**](https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset/data)

**Dataset 3**

[**https://www.kaggle.com/datasets/dev0914sharma/customer-clustering**](https://www.kaggle.com/datasets/dev0914sharma/customer-clustering)

**Clustering Algorithms**

**2**

**K-Means**

K-Means is a clustering algorithm that initially requires a parameter k which represents the number of clusters.

K-Means algorithm is the following:

Step 1: k initial centroids are randomly chosen across the dataset.

Repeat:

Step 2: each data instance is assigned a cluster by assigning it to the closest centroid.

Step 3: centroids are recomputed by finding the “center of mass” of each current cluster.

Until centroids do not change (or change insignificantly, for example too few instances change their assigned clusters).

Other stopping conditions might be used, like limiting the number of iterations. This could work well in many scenarios because almost all convergence happens during first few iterations. This method is also useful because of the fact that algorithm might stuck in an infinite loop.

Advantages of the K-Means are that it is easy to implement and easy to interpret results. It is computationally fast efficient as well. The disadvantage, however, is that the number of clusters is not known in the beginning, so it is not clear which value of parameter k to choose. Besides, as the initial centroids are chosen randomly, it leads to different clustering on the same dataset from one run of the algorithm to another. Moreover, K-Means does not work well when it comes to dataset containing non-spherical clusters.

**Hierarchical (Agglomerative)**

Hierarchical clustering

**BDSCAN**

**Results**

**2-3**

**Discussion**

2

**Conclusions**

0.5

**References**

Atasu, A., Duran, S., & Van Wassenhove, Luk N. (2022, June 18). The Dark Side of Solar Power. *Harvard Business Review.* <https://hbr.org/2021/06/the-dark-side-of-solar-power>

Bauer, S. (2020, December 2). *Explainer: The Opportunities and Challenges of the Lithium Industry.* <https://dialogochino.net/en/extractive-industries/38662-explainer-the-opportunities-and-challenges-of-the-lithium-industry/>

Cambridge University Press. (n.d.). Solar panel. In *Cambridge Dictionary*. Retrieved May 19, 2023 from <https://dictionary.cambridge.org/dictionary/english/solar-panel>

Crail, C., Tynan, C. (2023). *Everything You Need To Know About Solar Batteries*. <https://www.forbes.com/home-improvement/solar/what-is-a-solar-battery/>

Donohue, C. (2020). *How are solar panels made?* <https://cleannrg.com.au/how-are-solar-panels-made/>

# Earnshaw-Osler, M. (2023, February 24). *The Social and Environmental Impacts of Lithium Mining.* <https://borrumenergysolutions.ca/blogs/blog/the-social-and-environmental-impacts-of-lithium-mining#:~:text=The%20process%20of%20extracting%20lithium,to%20long%2Dterm%20ecological%20damage>

Mishra, A. (2015, June). Impact of silica mining on environment. *Journal of Geography and Regional Planning, 8*(6), 150-156. <https://academicjournals.org/journal/JGRP/article-full-text-pdf/915EC0C53587>