CONSTRAINED K-MEANS CLUSTERING ALGORITHMS IN PYTHON

Primary Design Report & Deliverables

Abstract

Machine Learning (ML) is a growing topic within Computer Science with applications in many fields. One open problem in ML is data separation, or data clustering. Our project is a validation study of, “Constrained k-means Clustering with Background Knowledge" by Wagstaff et. al. Our data validates the finding by Wagstaff et. al., which shows that a modified k-means clustering approach can outperform more general unsupervised learning algorithms when some domain information about the problem is available. Our data suggests that k-means clustering augmented with domain information can be a time efficient means for segmenting data sets. Our validation study focused on six classic data sets used by Wagstaff et. al. and does not consider the GPS data of the original study related applications. This material is based upon work supported by the National Aeronautics and Space Administration issued through the Oklahoma Space Grant Consortium.

Nicholas R. McDaniel

mcdanieln@student.swosu.edu

Table of Contents

Introduction

This document will cover what this project will entail. This project is one that the team has been working on for the past several months and will continue throughout the coming years. It has started as a validation study on Dr. Kiri Wagstaff's paper, "*Constrained k-means clustering with background knowledge*." That said, it is not constrained to this paper only. The project has blossomed through the several months of work and will continue to grow.

In the very first few weeks of the project, the biggest hurdle was tackling exactly what constrained *k*-means clustering *was* and how it was useful. This proved pretty simple to get through, as there is a plethora of guides online to explain exactly what *k*-means is, and the benefits to it.

After this, it transitioned to trying to get into the reproduction of the results found in the original paper. This also proved pretty easy, as Dr. Wagstaff provided her findings easily accessible on her website. It is at this point that we decided to start creating an application that will make this much more accessible to anyone who is interested in constrained *k*-means clustering, and a guide that will walk the inexperienced users step-by-step through the process that was taken in order to get to the final point.

We decided to turn to GitHub in order to provide this service. We will be hosting this repository on the public SWOSU repository, and it will be accessible freely to all. That was a simple choice. After this, it was the point that we needed to decide what language we needed to program the application in. Initially, R was looking favorable due to how happily it plays with data science. C# was also looking favorable due to our overall familiarity with the language. Ultimately, Python was pitched, and it stuck due to how lightweight it was and how easily it integrates with most operating systems. We reached out to Dr. Wagstaff and got her blessing, and then we proceeded. All of this considered, Python lacks the strength in GUI, so it was something unique to tackle that was not considered by the team initially. Additionally, the team was overall inexperienced with the scripting language, and had to learn the ins and outs in a sort of crash-course.

Deliverables

- Figure out what constrained *k*-means clustering is and how it is useful
    - *Appendix A* includes an example of what *k*-means is, and it is included in the repository as well
- Create a poster to deliver at both Oklahoma Research Day & SWOSU Research Fair
- Create a repository on SWOSU's GitHub page
    - This repository will include:
        - Guide to the Python program
        - Guide to explaining *k*-means
        - Relevant pages and notes
- Create a rudimentary GUI for the Python script
    - The GUI will be detailed in *Appendix B*
    - This GUI will do three things. This are:
        - Have an input box to enter how many clusters to make
        - Have a button to import an excel file
        - Have a button to take the input box and find the *k*-means of the excel file
- Report weekly findings to Dr. Jeremy Evert and Mrs. Madeline Baugher

Design

The design of this program is simple. It just takes an input, an excel file, and finds the *k*-means of it. For such a simple idea, it proved to be a difficulty on how to really provide this in an efficient manner. The application I ended up with was as simple as could be, allowing for future manipulation of the design as needed. The application will be able to also include a slider that will show the steps that were taken in order to reach the final result that is displayed. So, it would show the different clusters taken. The application will also be able to take in the much more common .data file and automatically convert it into a .xls file.

Appendix A

Figure 1:

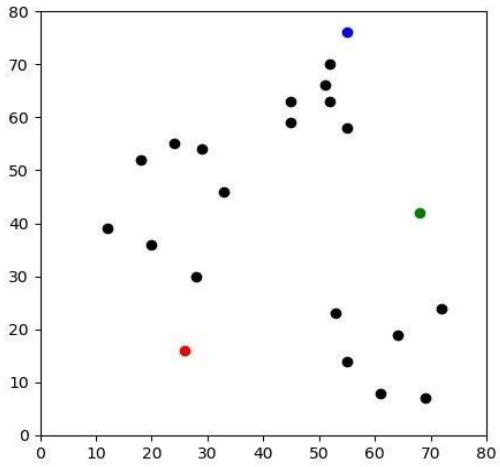


Figure 2:

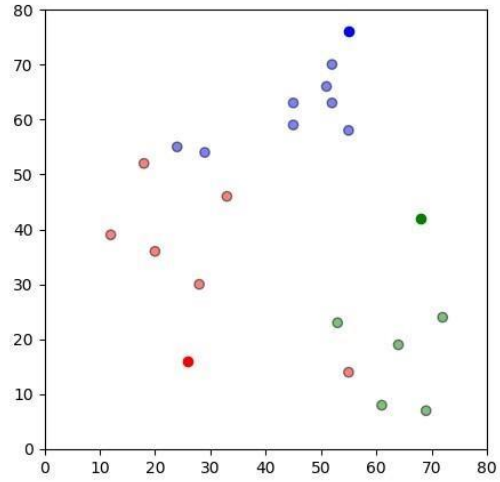


Figure 3:

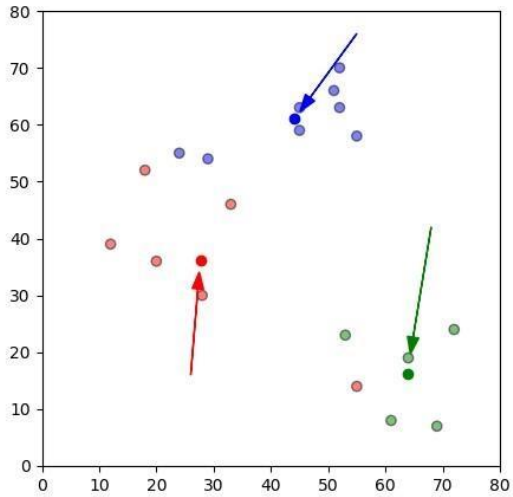


Figure 4:

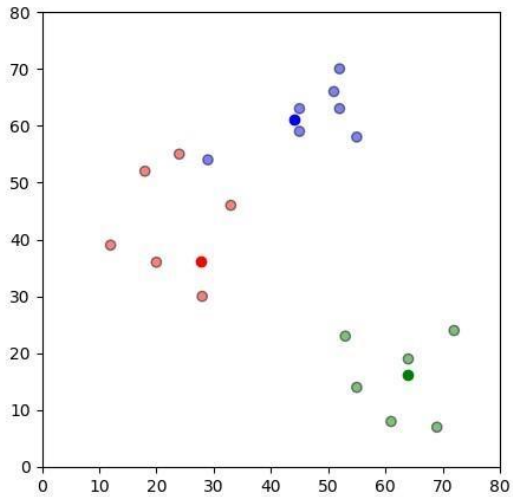


Figure 5:

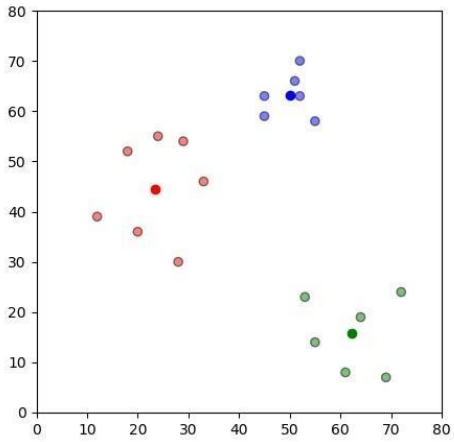
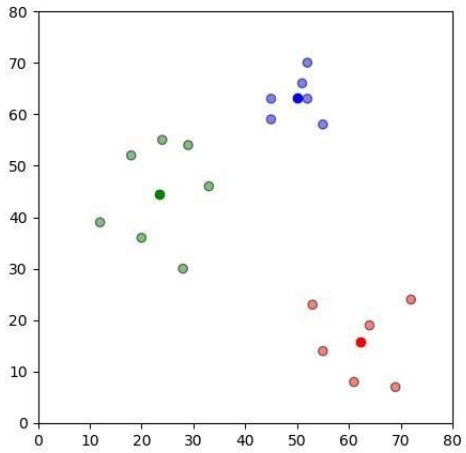


Figure 6:



Appendix B: Start of the Python File

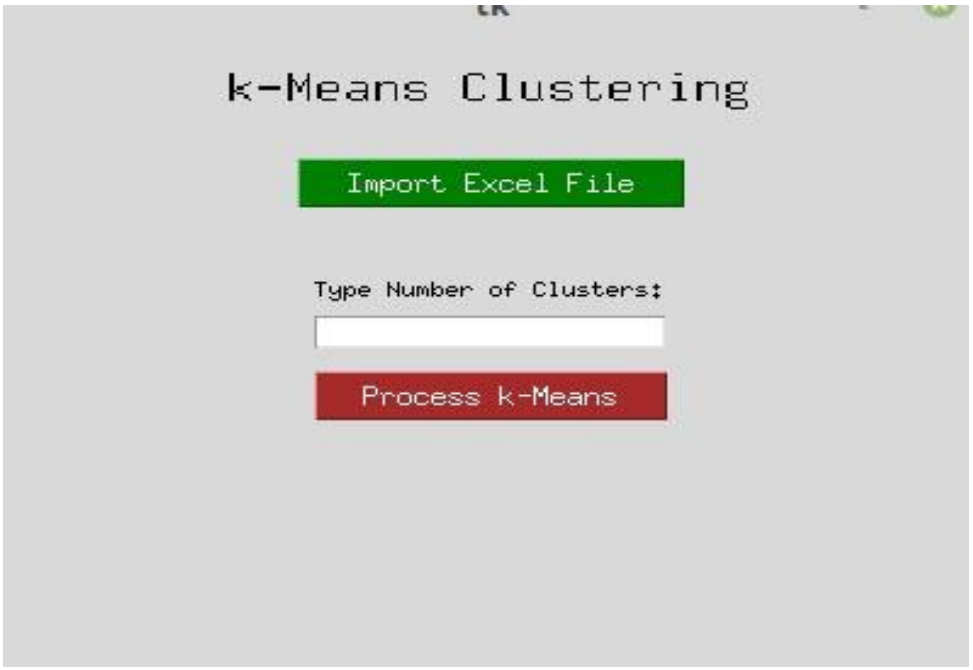
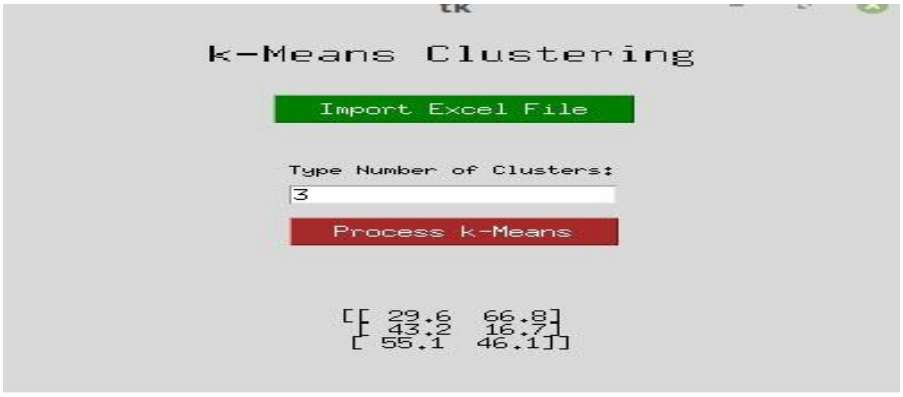


Figure 1. The start of the GUI



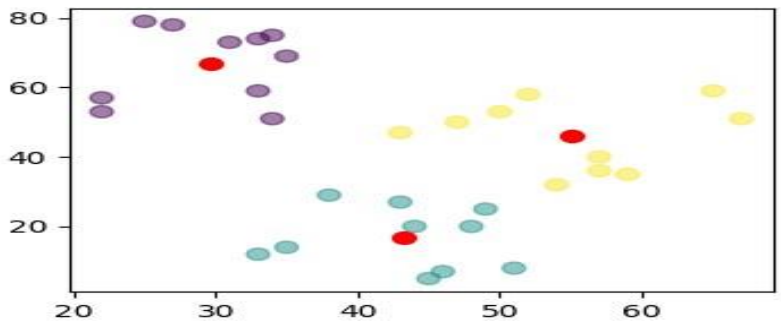
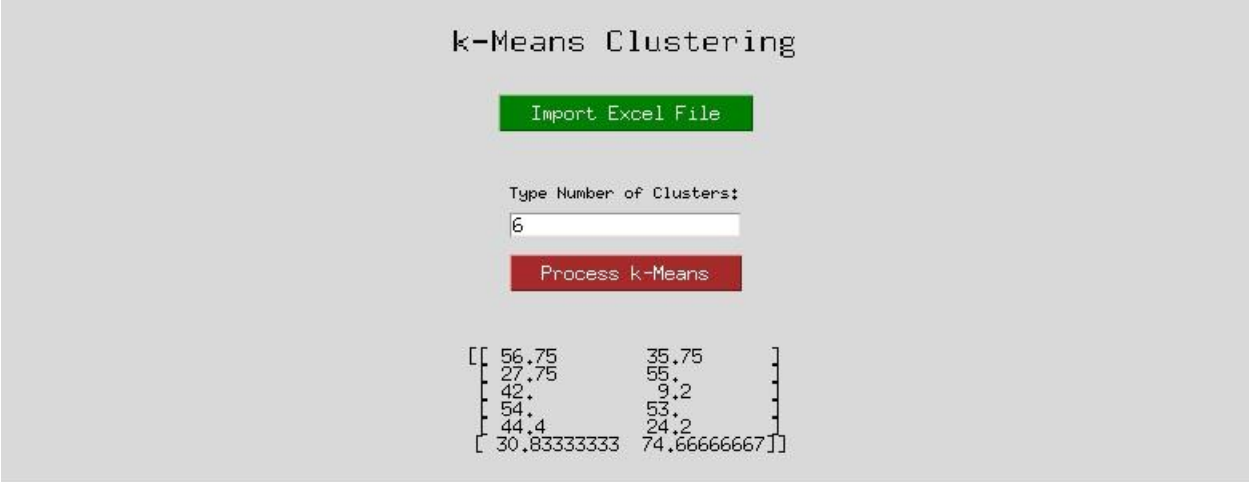


Figure 2. The first wave of clustering.



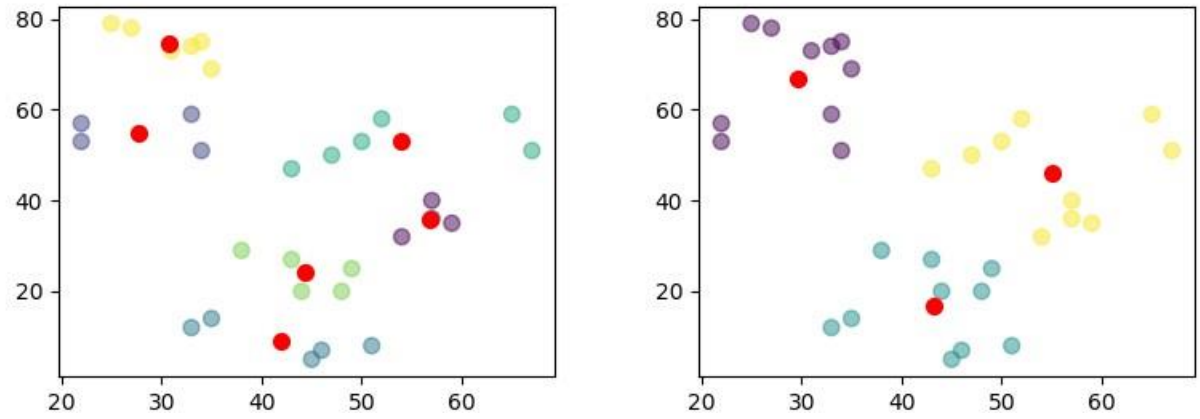


Figure 3. The example showing multiple different runs of clusters.

Works Cited

Basu, Sugato, Mikhail Bilenko, and Raymond J. Mooney. "A probabilistic framework for semi-supervised clustering." In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 59-68. ACM, 2004.

Bradley, P. S., K. P. Bennett, and Ayhan Demiriz. "Constrained k-means clustering." Microsoft Research, Redmond (2000): 1-8.

McDaniel, Nicholas; Burgess, Stephen; and Evert, Jeremy, "Constrained k-Means Clustering Validation Study" (2019). *Student Research*. 20.

https://dc.swosu.edu/cpgs_edsbt_bcs_student/20

McDaniel, Nicholas; Burgess, Stephen; and Evert, Jeremy, "Constrained K-Means Clustering Validation Study" (2018). *Student Research*. 12.

https://dc.swosu.edu/cpgs_edsbt_bcs_student/12

Wagstaff, Kiri, Claire Cardie, Seth Rogers, and Stefan Schrödl. "Constrained k-means clustering with background knowledge." In ICML, vol. 1, pp. 577-584. 2001.

Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks 16, no. 3 (2005): 645-678.