FINAL REPORT FOR OKLAHOMA NASA SPACE GRANT RESEARCHER NICK MCDANIEL

CONSTRAINED K-MEANS CLUSTERING ALGORITHMS

Abstract

Nick McDaniel was selected as a student researcher with the Oklahoma NASA Space Grant Consortium thanks to funding donated by Dr. Victoria Snowden. For this research, Nick provided a validation study of NASA JPL researcher Kiri Wagstaff. This research focused on Machine Learning.

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 under Grant No. NNX15AK02H.

Nicholas R. McDaniel
mcdanieln@student.swosu.edu

Table of Contents

Introduction

For the 2018-2019 school year, our research has been focused on a validation study of a machine learning project by NASA JPL researcher Kiri Wagstaff. 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. After a significant literature search, and a better understanding of machine learning, and data classification, the team had a better understanding of the topic. The fundamental concepts of *k*-means clustering are approachable thanks to a plethora of guides online to explain exactly what *k*-means is, and the benefits to it. But there is still a significant volume of terminology to wrestle with while putting a complete picture together.

After coming to grips with working definitions for terms related to *k*-means and machine learning, the team transitioned to the validation study. The goal of this effort was to recreate the results of the results found in the original paper. Thanks to the excellent writing by Dr. Wagstaff, the quality of the editing of their research, and the publication of their data along with their paper made this task approachable for the undergraduate team. The team developed an application. The goal is that our results will be much more accessible to anyone who is interested in constrained *k*-means clustering. It is our hope that the notebook will also serve as 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.

The team leveraged GitHub for source code management. The repository URL is https://github.com/swosu/constrained-k-means and is available to all members of the SWOSU Github team. The repository will be made public by the end of the semester, and it will be accessible freely to all.

The team then selected a language. 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. The visualization packages that worked well with our data sets were not as intuitive, and required a significant time investment. Handling the large sets of data required a significant level of scripting work. This was a new learning area for the team and also required a significant time investment.

Deliverables

The team established a set of deliverables to break the project up into manageable components. Here we discuss these deliverables and identify where the resulting documentation can be found.

- Define constrained *k*-means clustering and understand its applications.
  - *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
  - https://dc.swosu.edu/cgi/viewcontent.cgi?article=1011&context=cpgs_edsbt_bcs_student
- 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
  - https://github.com/swosu/constrained-k-means
- 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 a final report Mrs. Madeline Baugher
  - Weekly reports given in person to Dr. Evert.
  - This document serves as the final report for Ms. Baugher.

Design

The goal of the team was to maintain a clean design. The application accepts an input file, such as an excel file, then calculates the *k*-means of the data set. Maintaining a clean workflow and software stack required considerable effort. The end result is a relatively clean set of programs. It is our hope that the software design allows for future manipulation and maintenance as needed by future researchers. Future goals include the ability to incorporate a progress slider that will show the steps that were taken in order to reach the final result. A second goal is to allow inputs of other data files that are more common and convert them inside the application to the appropriate format for the existing software.

Results

We feel this was a successful project. The team was able to present the research at the SWOSU Research Fair and Oklahoma Research Day. The team was also able to post working Python code to the SWOSU repository. This research provides a good initial point for future researchers to build on. The paper concludes with a list of works cited. Example output images and screen captures of the GUI are listed in the appendices.

Oklahoma NASA Space Grant Spring 2019 Deliverables Report

Here we address our deliverables for the spring 2019 semester. This is for the research team member Nick McDaniel for the Dr. Snowden Memorial Scholarship under the NASA Oklahoma Space Grant Consortium

1. Nick did successfully submit 4 applications for NASA Internships at NASA centers for the fall 2018 semester. (4 is the limit the NASA internship system allows, one of which will be the Lucy mission internship.) Nick was also accepted to graduate school at Oklahoma Christian for their master's in computer science program.
2. Nick published two posters and posted a Github repository on his validation study of work by Kiri L. Wagstaff, a researcher with the Machine Learning and Instrument Autonomy Group at the Jet Propulsion Laboratory in Pasadena, CA.
3. Nick was a presenter at the 2019 Oklahoma Research Day and had his abstract published in the 2019 Oklahoma Research Day abstract book.
4. Nick did not finish publication of a guide for using Jupyter Hub for Computer Science 1 and/or High School students learning to program.
5. Nick completed his part of work with the NASA L'SPACE academy. Nick's team lead was not successful in submitting the team final report for the team's primary design report (PDR).
6. Nick did meet with Jeremy Evert for one hour a week and discuss his progress for both the Snowden NASA Internship and the L'SPACE academy. The L'SPACE academy had significant team member changes throughout the semester. The discussions focused on progress on the machine learning projects.
7. Nick emailed weekly reports to Jeremy Evert or had in person or Discord based meetings.

Nick McDaniel: ___________________________________
Jeremy Evert: ___________________________________
Madeline Baugher: ___________________________________

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.

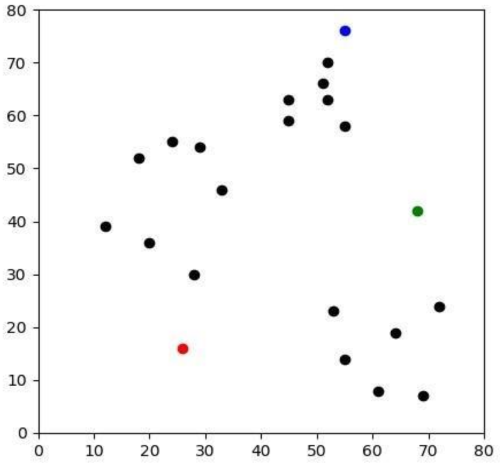
https://dc.swosu.edu/cpgs_edsbt_bcs_student/12

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.

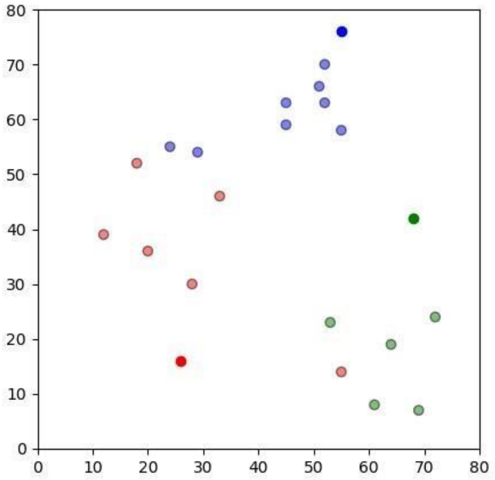
Appendix A

Figure 1:



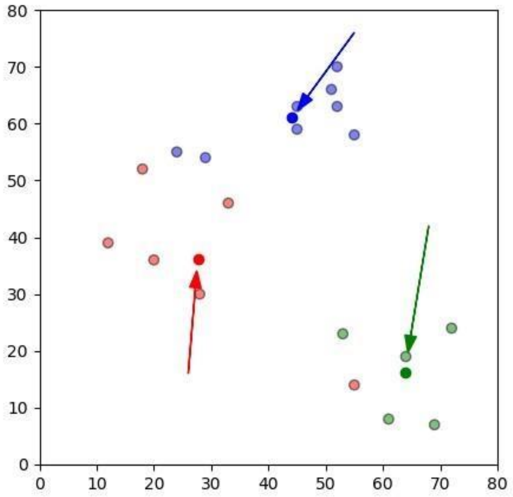
Initial cluster for an example test data set.

Figure 2:



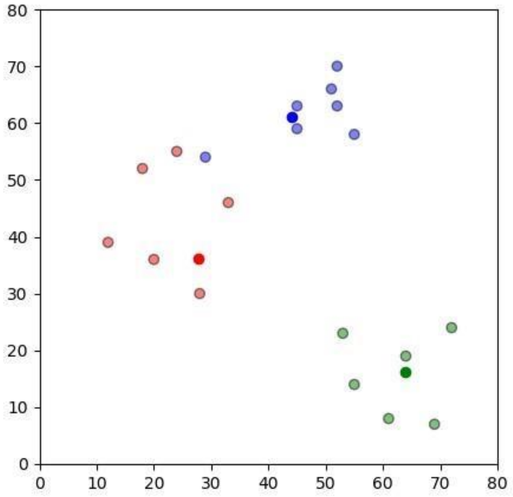
First assignment of centroids for the sample data set.

Figure 3:



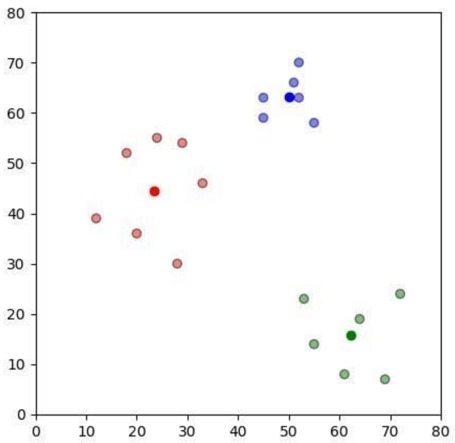
This shows the movement of centroids. The arrows point to how the centroids have moved from their origins to their new locations.

Figure 4:



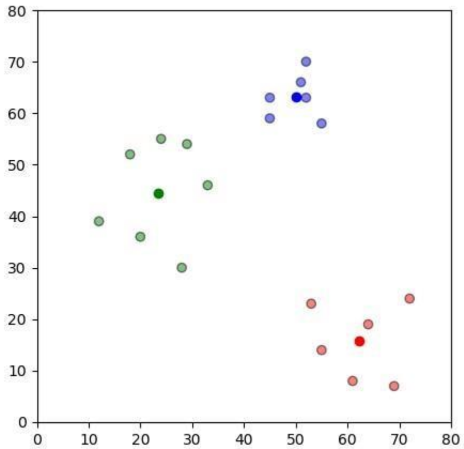
This image shows how the clusters are reassigned to the new centroids with their new locations.

Figure 5:



This shows where the new centroids are located based on the new classifications. This data set shows the centroids getting closer and closer to the center of the cluster it is assigned.

Figure 6:



This shows the final location of the centroids in the assigned clusters.

Appendix B: Start of the Python File

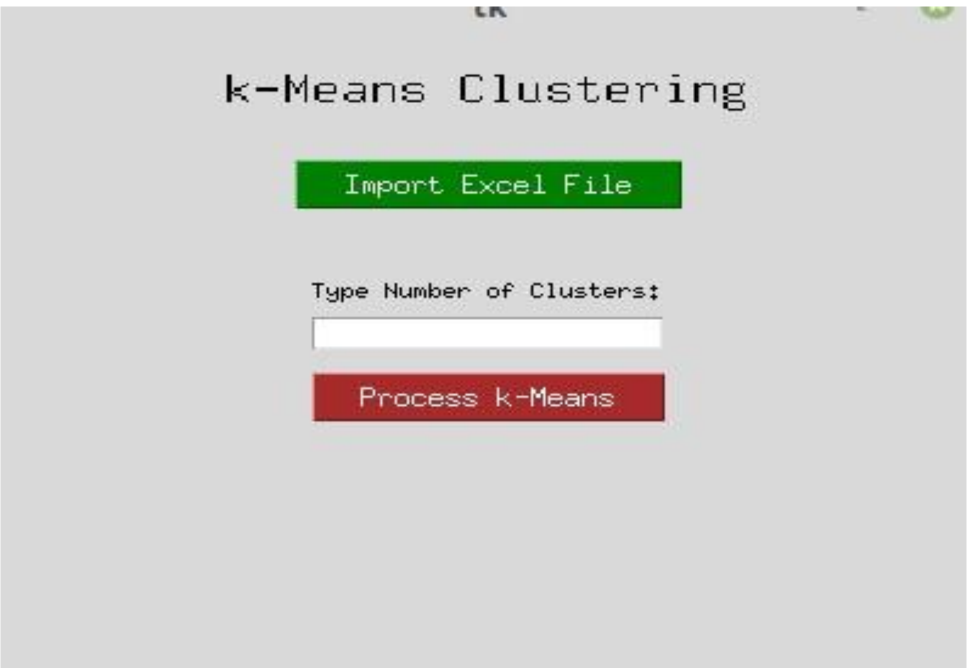
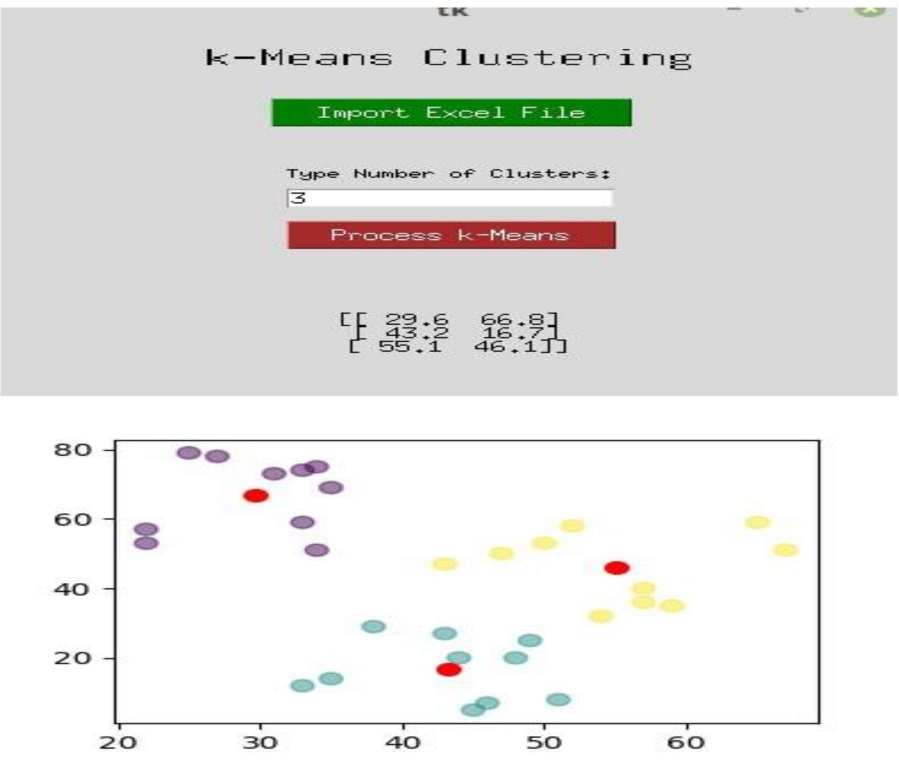


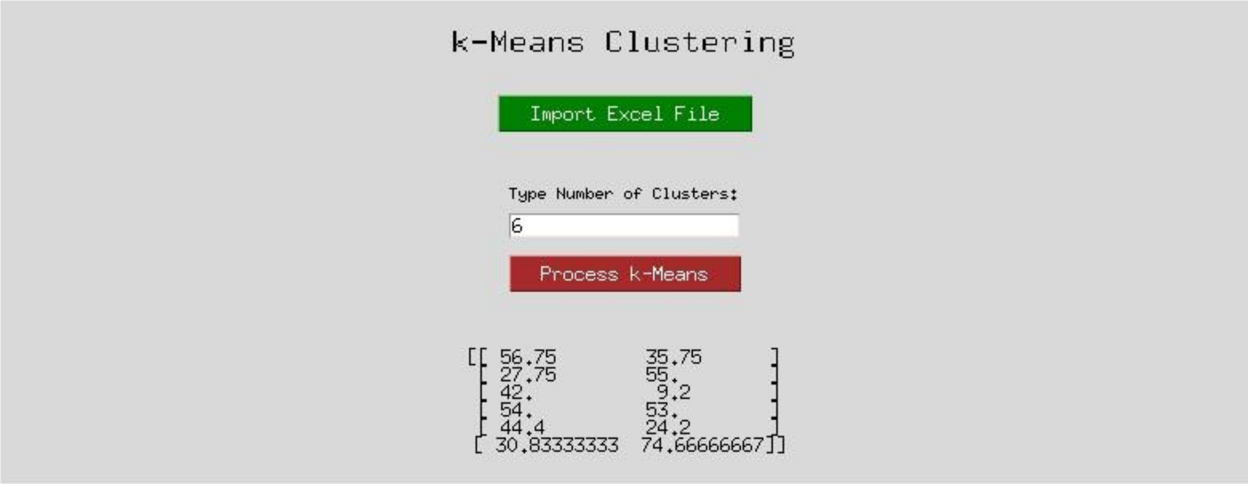
Figure 1. The start of the GUI



Type Number of Clusters: 3

```
[[ 29.6  66.8]
 [ 43.2  16.7]
 [ 55.1  46.1]]
```

Figure 2. The first wave of clustering.



Type Number of Clusters: 6

```
[[ 56.75        35.75      ]
 [ 27.75        55.       ]
 [ 42.           9.2      ]
 [ 54.          53.       ]
 [ 44.4         24.2      ]
 [ 30.83333333  74.66666667]]
```

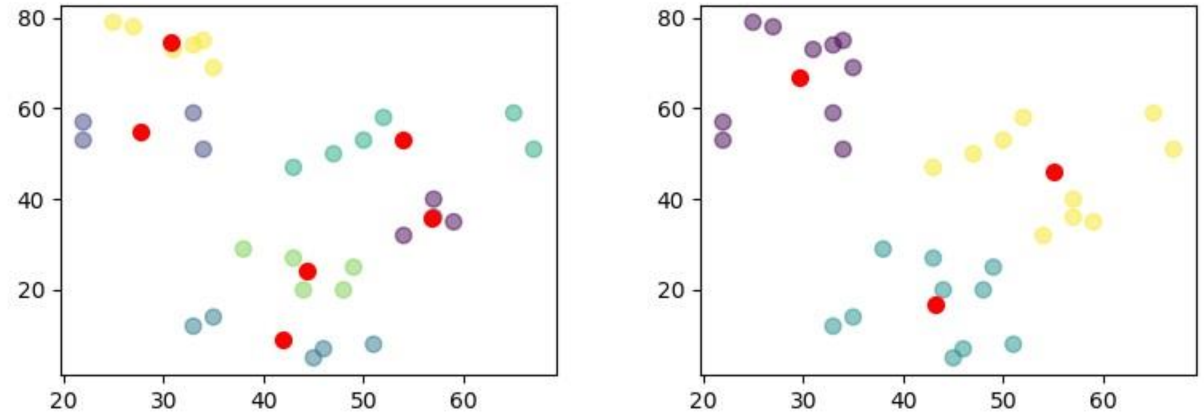


Figure 3. The example showing multiple different runs of clusters.