15.095 Project Proposal

*Names of team members and emails:*

Zachary Horton ([zhhorton@mit.edu](mailto:zhhorton@mit.edu)) and Tanner Street ([street24@mit.edu](mailto:street24@mit.edu))

*Problem Summary:*

Currently, K-Means Clustering is performed using a heuristic algorithm, with results relying on random initialization. We would like to develop a methodology which would provide an optimal cluster configuration using mixed-integer optimization. Primarily, our focus is to improve the reliability and flexibility on the existing algorithm with the use of modern-optimization solvers (i.e., Gurobi). Utilizing recent advancements in the world of linear and non-linear optimization, we believe that the K-Means can be optimized, allowing our team to gain a better understanding of how MIO can be applied to ML tasks, similar to methods discussed in class.

*Dataset(s) you plan to use (for applied projects):*

We plan on comparing several different metrics (computation time, # of iterations, average distance across clusters, etc.) of our method versus the typical heuristic results for several different datasets:

|  |  |  |
| --- | --- | --- |
| **Dataset** | **N (rows)** | **P (predictors)** |
| Abalone | 4177 | 9 |
| Auto MPG | 392 | 8 |
| Comp Hard | 209 | 7 |
| Concrete | 1030 | 8 |
| Housing | 506 | 13 |
| Space Shuttle | 23 | 4 |
| WPBC | 46 | 32 |

*Note: We may attempt different datasets as well, but this list came from Dr. Bertsimas’ lectures when comparing a heuristic algorithm to optimal MIO algorithms*

*Methods that you will use and how these relate to our class:*

* K-Means Clustering: unsupervised ML algorithm that we have not discussed in-depth in this course. Having a better understanding of this method, and especially unsupervised methods, would extend beyond the typical scope of the course and into new ML territory.
* MIO: mixed-integer optimization has been at the heart of many discussions this semester so naturally, applying MIO in a new setting would be a great way of getting a hands-on understanding of how we may apply and extend such topics to new problems.

*Challenges, and ideas to overcome them:*

Challenge #1: Reporting and Comparing Results

* It may be difficult to initially compare results from our algorithm to the heuristic one (from scikit-learn, or other libraries),
* Resolution: be very intentional with reporting results to allow for accurate comparisons

Challenge #2: Potential Non-Linearity

* If we decide to use Euclidean (L-2 norm) distance as our measure, it may be difficult to solve such problems using Gurobi, as opposed to just.
* Resolution: Begin to develop our algorithm using Manhattan (L-1 norm) distances, and extend capabilities as possible to other distance measures (L-2, L- norm, etc.)