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 (KMN) is performed using a heuristic algorithm, so the results rely on random initialization. We thus aim to develop a mixed-integer optimization (MIO) model that determines optimal cluster configurations. Implementing this model in a solver such as Gurobi will improve the stability and performance of the existing algorithm. Accomplishing this task will deepen our understanding on the application of optimization to machine learning tasks.

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

We plan on comparing several different metrics (computation time, number of iterations, average distance across clusters, etc.) of our method to 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: KMN is an unsupervised ML algorithm that we have not discussed in-depth in this course. Having a better understanding of this method (and, more generally, unsupervised methods) would extend our knowledge beyond the scope of the course into new ML territory.
* Mixed-integer optimization: MIO has been at the heart of many discussions this semester. Applying MIO in a new setting would be a great way to get novel hands-on experience with applying optimization methods to new machine learning tasks.

*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 and clear with reporting results to allow for accurate comparisons

Challenge #2: Potential Non-Linearity

* If we decide to use Euclidean (L-2 norm) distance instead of a linear measure, it may be difficult to find solutions using Gurobi,
* Resolution: Begin to develop our algorithm using Manhattan (L-1 norm) distances and extend capabilities to other distance measures (L-2, L- norm, etc.), if possible.