Math 5750/6880: Mathematics of Data Science
Project #2
project final report due October 2, 2025

A GitHub repo for this assignment is located here:

https://github.com/math-data-science-course/Project2

A LaTeX Project2 report template is available on the Canvas Project2 page. You should modify this template to produce your project final report in pdf format.

1. (Clustering Gaussian Blobs using k-means) In this exercise, you will perform a cluster analysis on a synthetically generated "Gaussian Blobs" dataset. Use the provided code in `Project2.ipynb` to import the dataset as a numpy array. Write code to perform a k-means cluster analysis with $k = 5$. Report your smallest k-means inertia value. Make a 2D visualization of your clusters via PCA, including both the clusters (colored by cluster) and the cluster centers. Additionally, make a confusion matrix that compares your assigned labels to the "true" labels. Here, you'll have to figure out how to best match the predicted and true labels. Explain your methodology in the report. Finally, perform an "elbow analysis" to justify the use of $k = 5$.

2. (Clustering Fashion-MNIST using k-means) In this exercise, you will import and perform a cluster analysis on the `Fashion-MNIST` dataset. Use the provided code to import the dataset as a numpy array. To get a sense of this dataset, first make a 5×2 array of figures, each plotting a distinct article of clothing. Write code to perform a k-means cluster analysis on this dataset, centering/scaling as appropriate. This is a larger dataset, so you may have to reduce the dimension/sample size as appropriate. In your report, explain your methodology and present your findings and figures.

3. (Dimensionality reduction for Fashion-MNIST) In this exercise, you will compare PCA and Random Projection on the `Fashion-MNIST` dataset. The goal of this exercise is to better understand how Random Projections performs as we vary dimension (Johnson–Lindenstrauss Lemma). Import the data and center/scale as appropriate. Implement both PCA and Random Projection methods for target dimensions $k \in \{10, 20, 50, 100, 200\}$. For each reduced dataset, compute the correlation between pairwise distances in the original standardized space and the reduced space. Make a plot of this correlation vs. k for the two methods. In your report, explain your methodology and present your findings and figures.

4. (Clustering Fashion-MNIST using spectral clustering) In this exercise, you will again perform a cluster analysis on the `Fashion-MNIST` dataset. Write code to use spectral clustering to cluster the data. Again, this is a larger dataset, so you may have to reduce the dimension/sample size as appropriate. In your report, explain your methodology and present your findings and figures. Compare your findings with Exercise 2.