

DATA 604, WINTER, 2021, Project 2

Due on February 9th, 2021, at 6:00PM.

1) (10 points) Use Matlab's `pca` function to find principal components of the dataset of handwritten digits collected by USPS. The principal components will be returned to you as a 256×256 matrix whose columns are the principal components. These 256 principal components form an orthonormal basis for the space \mathbb{R}^{256} and they are ordered in the descending order of component variance (think of it as from most important to least important). Plot the first and last principal components as functions on $[1, 256]$.

2) (15 points) Represent your data in the new basis of principal components (recall what it means from your Linear Algebra class). Reshape each representation as an image and plot the first of each handwritten digits class (1, 2, ..., 9, 0) as 16×16 images.

3) (15 points) Treat the `pca` representation of USPS digits as your new data set. Perform the kNN classification on this new data using your optimal parameters from Problem 2 on Project 1. Report the success rates for each digits and the overall accuracy.

4) (10 points) Compare your results with those obtained on Problem 2 in Project 1 and discuss the outcomes. Does `pca` help or hurt your classification?