MATH 420, HW 6 (Machine Learning Project 1 HW 1), SPRING 2015, Due: March 24, 2015

This HW is about the role of representation methods in classification and clustering problems.

As in HW # 5 we shall represent each of the images as a vector with 256 coefficients with 2 basic schemes: reading the images row by row, from left to right, and reading the images column by column, from top to bottom.

For each of the two vectorization schemes, plot the whole USPS data set (training and testing together) by using a pair of coordinates. Start with something simple, like 1st and 2nd; or 255th and 256th; or a randomly chosen pair. Use different colors of points for different digit classes in your visualization. Are different digit classes separated from each other?

Next try to find a pair of coefficients that might provide best visualization. This is more akin to classical modeling: think of what the data is and interpret it as best as you can. Do you see an improvement in visualizations? Do you have to change your ”best” pair of coordinates when switching from one vectorization model to the other? Are the plots for different vectorization schemes similar or different?

Next, perform PCA and plot some pairs of most significant principal component expansions. Do we now see any improvement in separation of digits? Is there any significant difference between two vectorization schemes?