Connie Okasaki
AMATH 584
Assignment 6
14 Dec 2020

See code with plots at github.com/cokasaki/amath584

(a) I used regular least-squares, lasso (ℓ_1 regularization), ridge regression (ℓ_2 regularization), and elastic net regression (both ℓ_1 and ℓ_2).

(b) I ranked pixels as follows: first since lasso is intended to promote sparsity I used this as the most important heuristic. I ran several lasso regressions with varying regularization parameters, and partially-ordered the pixels based on the highest parameter at which the pixel had any non-zero across all 10 digits. I then ordered the pixels within each level according to the sum of absolute values of the coefficients across all 10 digits in the regression at which the pixel entered.

(c) The "elbow" of the MSE curve occurred at about 75/784 pixels (same pixels for each digit). Using less than 10% of the pixels we were able to achieve an MSE only 20% higher than using all the pixels. Alternatively, using half the pixels we could have achieved nearly the optimal MSE. This is due to about half the pixels lying around the edges where no digit actually reaches. Although classification using a naive linear regression model is a little dicey, using the heuristic that the maximum predicted digit score should be the classifier output we were able to achieve an 80% accuracy rate.

(d) Redoing the analysis we found that the optimal pixels for each digit varies substantially from digit to digit. Approximately ≤ 50 pixels are necessary for each digit. Some of these, particularly "4," "7," "8," and "0" clearly correspond to an outline of the digit but others are less interpretable.

(e) Solving $AX = B$ corresponds to a linear regression view of this problem. However, for classification problems the output of a linear regression is unclear. What does a score outside of the range $[0, 1]$ mean? A more common approach is logistic regression, in which the output can be interpreted as a probability. For more than one class, a common approach to extend logistic regression is the one-vs-all classifier, which again corresponds to this situation, where we have defined each B to be a unit vector in the digit-th direction. Using logistic regression instead of linear regression we achieve a slightly lower accuracy rate, but with the advantage of much more interpretable results. Likely better choices of regularization parameters by use of a validation set could have improved these accuracy rates.