

Proposal for Dimension Reduction

1 Methodology

We had two ideas for dimension reduction of our variables. One was averaging the pixel intensities of the 28x28 image to create a 7x7 with less pixels. Another was simply taking out the pixels that were completely correlated with one another (all intensities were zero for that one pixel for each image in training set). These two methods illustrated above are the "visual" was of thinking about dimensional reduction. Both are a type of PCA, where we essentially take linear combinations of the pixel intensities. However, we also plan to run PCA using R and determine the "mathematical" best way to reduce our dimensions.

2 Pixel Intensity Averaging

3 Removing Collinear Pixels

Removing collinear pixels was pretty simple because we had accomplished this once already when doing lda and qda. Essentially, we are taking a linear combination of the pixels, with each weight being a zero or a one based on whether we decide to remove it or not. In doing this dimension reduction, we are not actually running PCA on the dataset, but rather taking out columns which do not contribute to the overall image as much according to us. Therefore, if the average pixel intensity of a specific pixel over the 59999 training images is less than 10, we removed the pixel from the reduced training set.

We reduced our number of variables down from 784 to 370, which is a significant decrease. The runtime for a test set of 200 images accordingly decreased from 49 seconds to 23 seconds. Interesting enough, we expected an increase in our test error since we included less explanatory variables. However, our test error actually decreased, from 0.04975124 to 0.04477612.

4 Mathematical Dimensional Reduction

5 Conclusion