STATS 760 A Survey of Modern Applied Statistics

Assignment 4 2017

Due Wednesday May 30. Email to Alan by 5pm.

This assignment uses a data set similar to the zip code data discussed in class. The data file train.csv contains 42000 gray-scale images of hand-drawn digits, from zero through nine.

The data are on the Kaggle website. To quote from the website:

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.

The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image.

Each pixel column in the training set has a name like pixelx, where x is an integer between 0 and 783, inclusive. Visually, if we omit the "pixel" prefix, the pixels make up the image like this:

000 001 002 003 ... 026 027
028 029 030 031 ... 054 055
056 057 058 059 ... 082 083
| | | | ... | |
728 729 730 731 ... 754 755
756 757 758 759 ... 782 783

Requirements.

1. Read the data into an R data frame. Check that all variables (except the first) are integers in the range [0:255] and that the values of the response (column 1) are in the set 0,1,….8,9. Change the response into a factor. [5 marks]

2. For the first 25 lines of data, draw the pixel array of each image, and label it with the actual digit. Lay out your 25 images in a 5x5 array on a single page. Some R code to help you do this is described at the end of this assignment. [5 marks]

3. Using the first 21,000 lines of data, fit a classification model using (a) random forests, and (b) support vector machines. For the latter, you will have to figure out how to turn the SVM (which is a binary classifier) into a classifier that will work for 10 categories. Calculate the PE for each model, using just the first 21000 lines of data. [20 marks]

4. Then, using the second set of 21000 digits (i.e. lines 21001 – 42000) as a test set, calculate a test set estimate of the PE for each method. Compare with the estimates you obtained in part 3. [10 marks]

Note: this data set is quite large. If you are having trouble getting the code to run, you could try the following (a) use a smaller training set, or (b) eliminate the variables that are “almost constant”. See the book *Applied Predictive Modelling* for details of the latter.

Code to draw the digit in the first line of the data file:

```
z = matrix(unlist(train.df[1,-1]), 28,28)
zz = z
for(j in 28:1)zz[,j]=z[,29-j]
image(zz, col =  gray((32:0)/32))
box()
```