Logistic Regression Homework (2/9/18)

Image Classification

In this problem, we'll tackle the task of image classification: labelling an image according to the content it represents. We'll be using a popular dataset of images of hand-written digits called the MNIST dataset. This dataset contains thousands of images of hand-written examples of single digits.

This dataset has been used as a test set for machine learning for a long time. For more details and the latest results, see e.g. http://yann.lecun.com/exdb/mnist/ and https://en.wikipedia.org/wiki/MNIST_database .

Preparation

Load the workspace **mnist_all.RData**, available in Canvas, into your RStudio workspace. There are two lists `train` and `test` in this workspace. Each contains three members, $x, $y, $n.

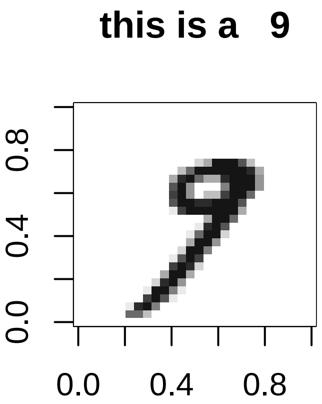
- `train$x` contains 60,000 rows and 784 columns. Each row represents a single image of a handwritten digit between 0 and 9 that was a 28x28 grayscale pixel image (entries in $\{0, \ldots, 255\}$ with 0 representing white and 255 representing black). The 28×28 matrix representing the image of a digit has been flattened into a 1x784 vector.
- `train$y` is a vector with 60,000 entries, each giving the digit in the corresponding row of `train$x`.
- `train$n` = 60,000, the number of entries in the training set.
- `test` is structured the same way, with 10,000 test observations.

You can the following function to "visualize" these images as R plots. *This function only works for the training data, but you can easily rewrite it so it works also for the test data.*

```r
plot_digit <- function(j){
arr784 <- as.numeric(train$x[j,])
col=gray(12:1/12)
image(matrix(arr784, nrow=28)[,28:1], col=col,
      main = paste("this is a  ",train$y[j]))
}
```

Example code:

```r
plot_digit(34)
```



Here are some recommended steps to familiarize yourself.

This does not have to be turned in.

- Plot examples of all 10 digits.

Each variable is the grayscale value of a pixel in an image, i.e. a "feature". Here are some ways to explore these features.

- Find several different features that have zero variability (all images have the same value in this pixel).
- Find several different features that each have positive variability.
- Pick several pairs of features that both have non-zero variability. Make a scatterplot of these two features against each other. Label the datapoints with colors corresponding to the digits.
- Make side-by-side box plots of the grayscale values for all 10 digits. *You will have to combine $x and $y into a single data frame.*

Homework Problems

In all three problems, build a classifier that distinguishes 0 from 1, using all training data for these two digits.

Problem 1

Build a classifier, using only one variable (pixel). It should have large variation. Give the summary of the model and write out the logistic regression equation that has been obtained. Determine the fraction of true positives, if the fraction of false positives is 0.1.

Problem 2

Build a classifier, using two variables that have small correlation. Find the Area under the Curve (auc) using the training data. Is this a good classifier? Make a scatter plot of the two variables, colored by the type of digit, and use this to explain the performance of the classifier.

Problem 3

Build a classifier, using the 10 variables with the largest variances. Make a ROC curve and comment on the performance of the classifier.