Logistic Regression Homework (2/10/17)

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. For simplicity, we'll focus just on a subset of the images: the 0's and the 1's.

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

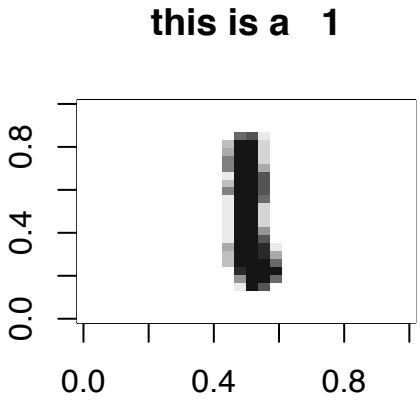
Load the dataset **mnist_data.RData**, available in Canvas, into your RStudio workspace. There is a single data frame **images_df** in this dataset. It has 2115 rows and 785 columns. Each row of the data frame represents a single image of a handwritten digit (0 or 1) that was a 28x28 grayscale pixel image (entries in $\{0, \ldots, 255\}$ with 0 representing white and 255 representing black). The 28×28 matrix representing the image has been flattened into a 1x784 vector, which is stored in the first 784 entries of the row. The 785th entry of each row is called **labels** and contains a 0 or a 1, depending on what the image in this row shows.

Use the following function to "visualize" these images as R plots. Examine a few of the rows of the matrix and get a sense of what these hand-written digits look like to the human eye. Make sure to see examples of each class.

```
plot_digit <- function(j){
arr784 <- as.numeric(images_df[j,1:784])
col=gray(12:1/12)
image(matrix(arr784, nrow=28)[,28:1], col=col,
      main = paste("this is a  ",images_df$labels[j]))
}
```

Example code:

```
plot_digit(34)
```



A.

Plot six different images, three 0's and three 1's.

B.

Each pixel of an image is a "feature", represented as a lightness-darkness value. Do some exploration of these features.

(i) Find five different features that have zero variability (all images have the same value in this pixel).

(ii) Find five different features that each have positive variability.

(iii) Pick any two features that have some non-zero variability. Make a scatterplot of these two features against each other. Label the datapoints with color corresponding to the **labels** column.

C.

Choose three different pairs of features that seem like they would allow separation of the two classes, then build logistic regression models for image classification that use these three pairs. Comment on the results and how well the models do. Provide confusion matrices for each model and give interpretations for them. Also compute AUCs, using the R package **pROC.**

D.

Build a logistic model that uses at least five features. Make a confusion matrix and compute the AUC. The result should be better than the best model that you found in **C**.

E.

Now try making a model with all the features. *This will probably take R several minutes and will result in some errors and warnings.* Look at the estimated coefficients and standard errors. Describe what you see. Explain why this approach fails.