In this assignment, you will be using logistic regression in the context of optical character recognition. Each line of the data set is a digital representation of a scanned handwritten digit, originally a part of a zip code handwritten on a US letter.

Each hand-written digit is represented as a 16 x 16 array of pixels, as in the examples below.

Each pixel is given a grey-scale value in the range $[-1, 1]$ with -1 representing white and 1 representing black. There are thus $16 \times 16 = 256$ numbers representing a particular digit, which we can take as the values of 256 variables, $V_1, \ldots, V_{256}$, say. The relationship between the pixels and the variables is as follows:

In this assignment, we will use the data on the digits 3 and 7 only. The data set in the file `train.txt` on the web page contains data on 1303 handwritten digits, 658 of which are 3’s and the rest 7’s. The file has no headers but the actual digit is recorded in the first column and the greyscale values in the next 256 columns (Variables $V_1$-$V_{256}$).

The aim of the assignment is to construct a prediction rule that will allow us to discriminate between 3’s and 7’s. Such a rule could be built into a letter sorting machine that scans the zip codes, converts each digit into a pixel array, and then evaluates the predictor.
1. Read the data into an R data frame, creating a variable \( D \) to represent the actual digit, and variables \( V1-V256 \) for the grayscale values. Check that all variables are in the range \([-1,1]\). Print out the first line of the data frame. [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. Study your array of 25 images. Which pixels (i.e. which variables) seem to be the most important for discriminating between 3’s and 7’s? [5 marks]

4. Calculate the correlations between \( D \) and the other variables. Identify the 20 most highly correlated variables. Do these correspond to the ones you identified in Question 2? [5 marks]

5. Fit a logistic model to the data, using the 20 variables you identified in Question 3. (The regression will not converge if you use all 256.) Calculate the fitted logits for each of the 1303 hand-written digits. [5 marks]

6. Use your fitted model to predict if a digit is a 3 or a 7 on the basis of the 20 variables. (Predict a digit to be a 7 if the fitted probability of a 7 is more than 0.5 (or equivalently, if the corresponding logit is positive.) Evaluate the “in-sample” prediction error (using the same data set to fit the model and evaluate the error). [5 marks]

7. Use stepwise variable selection to chose a submodel of the 20-variable model. What are the in-sample prediction errors for this model? [5 marks]

8. Use the data set \texttt{test.txt} to calculate an estimate of the “out-of-sample” prediction error for the chosen models. Here, we use the original data set to fit the model, and the new data to calculate the prediction error. Also, use cross-validation to estimate the prediction error. Comment on any differences between these errors. [5 marks]

**NB: you must include your R code in your answers.**

**R code for drawing the digits**

The following code will take a vector \( v \) containing the values of \( V1--V256 \) and a variable \( d \) and draw the pixel array, labelling it with the actual digit \( d \).

```r
z = matrix(unlist(v), 16,16)
zz = z
for(j in 16:1)zz[,j]=z[,17-j]
image(zz, col = gray((32:0)/32))
box()
text(0.1,0.9,d, cex=1.5)
```