Lab 4 - Classification

Yuval Benjamini

Due Tuesday June 30th before 4:30pm

Hand In Procedure: Labs can be handed alone or in pairs (no more than 2 per lab!). Please prepare a file with a writeup and code (the writeup can be in Hebrew or English). Please make sure the r

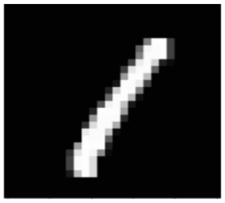
1 Classification Lab

We will try to classify handwritten digits in $28x28$ greyscale values by their digit. The images are from the MNIST dataset; you can (and should) read more about the dataset here: http://yann.lecun.com/exdb/mnist/.

For your convenience, I have supplied a script (written by Prof. David Dalpiaz from UIUC) that downloads the data and prepares it for R. `load_MNIST.R`

Your goal is to build a classifier for telling apart the digit 1 from the digit 7.

1. I'd like you to compare two methods, each from a different classifier family (e.g. Classification trees, Bayesian classifiers, linear discriminants, etc). You can use as features the raw pixel values (0 for white to 255 for black). [1] Explain your decisions in designing the classifier, specifying the penalty and any hyper-parameters, and in choosing the threshold.
2. Write your own function that calculates a confusion matrix for both the training and the testing sets.
3. Write your own function that draws a response operating curve (ROC). Draw ROCs for both classifiers.
4. For one of your classifiers, display four examples that were classified *incorrectly.* Can you see what made these examples hard for the classifier?
5. Here is an image of a white digit (the digit 1) on a dark background. Do you expect both of your fitted classifiers to work well on this image? Why or why not?



   [Hint: Think how would this image be coded into numbers? what would happen if you try to classify using your method?]

[1] Please do not use online tutorials for this data (MNIST) or code from other courses.