Machine Learning Module

Week 4

Laboratory Exercise, Week 5

Non-Probabilistic Classification Methods

Mark Girolami
girolami@dcs.gla.ac.uk
Department of Computing Science
University of Glasgow

January 29, 2006

1 Laboratory Exercise

There is one data set available for download from the class website for this laboratory named `digits_3_8.mat` and is a binarised collection of the handwritten digits 3 and 8. Some examples are given in Figure (1).

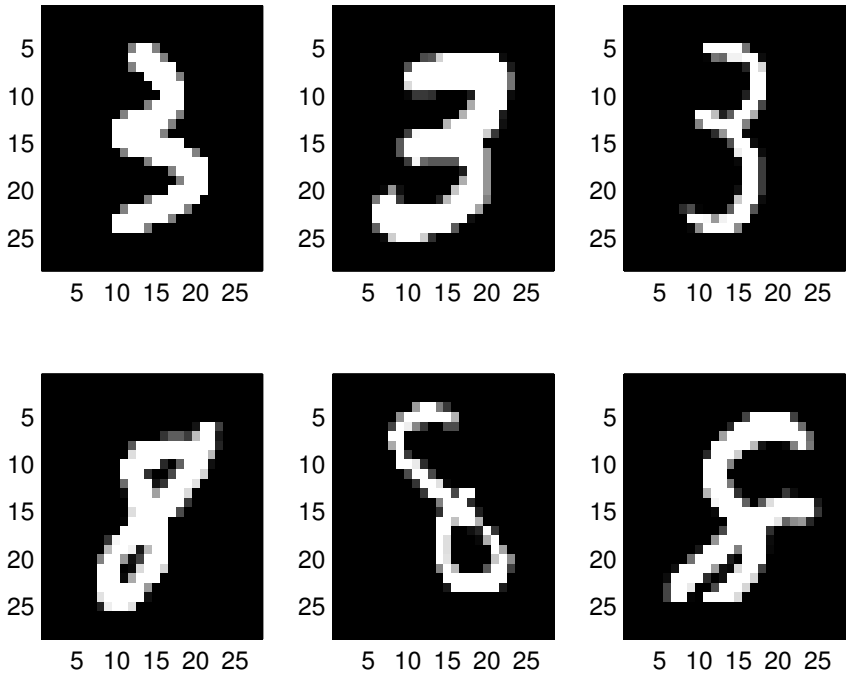

Figure 1: The top row gives examples of the digit 3 whilst the bottom row give examples of the handwritten digit 8. Each of the images are 28×28 pixels in size, with each pixel taking on 8 bit values.

The matlab file has a 400×784 matrix $\mathbf{X}$ which contains 200 examples of the digit 3, and 200 examples of the digit 8. The 28×28 pixel images have been reshaped into 784×1 dimensional vectors, hence the 400×784 matrix $\mathbf{X}$, which will be used for training purposes. There is a corresponding label vector $\mathbf{t}$ which encodes the two classes as ± 1. The images to be used for testing are stored in the matrix $\mathbf{Xt}$ and the corresponding labels being found in the vectors $\mathbf{tt}$. There are 500 examples of each digit available for

testing classifier performance.

1. Download this Matlab file and view the data using for example `imagesc(reshape(X(241,:),sqrt(784),sqrt(784))')`. Look up the Matlab Help files to study the `reshape` and `imagesc` commands.

2. Use the KNN classifier and study how the test error varies with the number of nearest neighbours used. Note that the target values for the KNN should use a 1,2 encoding rather than +1, -1 code. Report the best performance achieved and plot how test error varies with K. What effect does standardising the digits data have on the KNN classifier performance?.

3. Use the SVM classifier and study how the achievable test error varies with the box constraint parameter C and the parameters of the kernel you decide to use. Report the best performance achieved and show plots of the test errors against C and any kernel parameters.

4. Visually examine the images of the digits which are support vectors and report on any interesting characteristics you observe about the *support digits*.