

## ASSIGNMENT 4, due October 27.

Homeworks are due in class or in my mail box by 2pm on the due date. *LFD* is the class textbook

### 1. (200) Handwritten Digits Data - Obtaining Features

You can download the two data files with handwritten digits data: training data (`ZipDigits.train`) and test data (`ZipDigits.test`). Each row is a data example. The first entry is the digit, and the next 256 are grayscale values between -1 and 1. 256 pixels corresponds to a  $16 \times 16$  image. For this problem, we will only use the 1 and 5 digits, so remove the other digits from your training and test examples.

- (a) Familiarize yourself with the data by giving a plot of two of the digit images.
- (b) Develop two *features* to measure properties of the image that would be useful in distinguishing between 1 and 5. You may use symmetry and average intensity (as discussed in class) or anything else you think will work better. Give the mathematical definition of your two features.
- (c) As in the text, give a 2-D scatter plot of your features: for each data example, plot the two features with a red 'x' if it is a 5 and a blue 'o' if it is a 1.

### 2. (400) Classifying Handwritten Digits: 1 vs. 5

Pick one of the following 3 classification algorithms for non-separable data:

- (i) Linear Regression for classification followed by pocket for improvement.
- (ii) Linear Programming for classification.
- (iii) Logistic regression for classification using gradient descent.

Use your chosen algorithm to find the best separator you can *using the training data only* (use your 2-D features as the inputs). The output is +1 if the example is a 1 and -1 for a 5.

- (a) Give separate plots of the training and test data, together with the separators.
- (b) Compute  $E_{\text{in}}$  on your training data and  $E_{\text{test}}$ , the test error on the test data.
- (c) Obtain a bound on the true out-of-sample error. You should get two bounds, one based on  $E_{\text{in}}$  and one based on  $E_{\text{test}}$ . Use a tolerance  $\delta = 0.05$ .  
Which is the better bound?
- (d) Now repeat using a 3rd order polynomial transform.
- (e) As your final deliverable to a customer, would you use the linear model with or without the 3rd order polynomial transform? Explain.