CSE 5523. Homework 2. Due March 26th in class.

Problem 1. Run a linear SVM on the two class dataset given online (you can use a standard toolbox). Compare its performance to that of the least squares linear classifier. How does the choice of the parameter C affect performance?

Instructions: download *79.mat*, which contains images of digits. Each image is given as a 28x28 matrix of grayscale pixel values. It is stored as a 784 ($= 28 \times 28$) array. You are given 1000 images of 7 and 1000 images of 9. These are stored as a single 2000×784 matrix in the file 79.mat. The first 1000 digits are sevens, the rest are nines. Download that file and type "load 79.mat" in Matlab. The matrix d79 contains the data. You can visualize the digits by typing, e.g., the following:

colormap(gray);

x = reshape (d79(1234,:),28,28);

y = x(:,28:-1:1);

pcolor(y)

This bit of code shows you the digit number 1234 (which is a 9).

Problem 2.

Implement (do not use standard toolboxes) the Least Squares classifier using gradient descent. Compare your results to standard least squares classifier (obtained using pseudo-inverse).

Problem 3. Reduce the dimension of the dataset (both train and test) to 400 using the Principal Components Analysis (we have not discussed it yet but you can use a standard toolbox). Apply linear regression and SVM (using large value of the parameter C) to 50, 100,150,...2000 training examples (i.e., $25, 50, \ldots, 1000$ from each class, you can choose them at random). Plot the error on the test set. Observations?

Problem 4. Use gradient descent (instead of the explicit solution) for linear regression in Problem 3. For 50, 200, 400, 1000 and 2000 training examples plot the dependence of the test error on the number of iterations. What do you observe?