CS340 Winter 2010: HW 1
Out Wed. 12nd January, due Frid. 21st January

1. Handwritten digit recognition

We consider the problem of handwritten digit recognition. You will use a subset of the standard MNIST database (http://yann.lecun.com/exdb/mnist/). We will use a subset of it (mnist_HW1.mat) which can be downloaded on the webpage of the course.

In this database, each training or test example is a 28-by-28 grayscale image. These images have been converted to vectors $28 \times 28 = 784$ by sorting the pixels in raster scan (row-by-row) order. The Matlab reshape command can be used to convert these vectors back to images for visualization.

For example, we can plot the fourth test example of class 2 as follows:

>> `imagesc(reshape(test2(4,:), 28, 28)');`

To reduce computational complexity and simulation time, in the following questions we focus on only three of the ten handwritten digits: "1", "2", and "7". In this question, we explore the performance of K-NN classifiers at distinguishing handwritten digits. We will restrict ourselves to the Euclidean distance.

1. Implement and submit a Matlab function which finds the K-nearest neighbors of any given test data, and classifies them according to a majority vote of their class. This function should be of the form

   `[ytest,etrain]=knn(xtrain, ytrain, xtest, k)`; where `xtrain` are the features of the training data, `ytrain` are the labels/outputs of the training data, `xtest` are the features of the test data, `k` is the number of neighbors in the K-NN classifier, `ytest` are the predicted output of the test data and `etrain` is the error rate on the training set.

   Using the given training data (which includes 200 examples of each class, i.e. $N = 600$ training data), what is the empirical accuracy (fraction of data classified correctly) of 1-NN and 3-NN classifiers on the given 600 test examples from these classes?

2. Consider in this question the 1-NN classifier.

   a) How many 1s are missclassified as being 2s, and how many as 7s? How many are correctly classified?

b) How many 2s are missclassified as being 1s, and how many as 7s? How many are correctly classified?

c) How many 7s are missclassified as being 1s, and how many as 2s? How many are correctly classified?

For each of the cases above, plot a few examples of correctly and incorrectly classified data. Do you see any patterns?

3. Implement and submit a function which uses 5-fold cross-validation, on the training dataset from part 1, to estimate the accuracy of a K-NN classifier. Determine a cross-validation accuracy estimate for eight candidate classifiers, produced by the use of $K = \{1, 3, 5, 7, 9, 11, 13, 15\}$ nearest neighbors. Plot of these accuracy estimates versus K. Which classifier is estimated to be most accurate?

4. For the classifier estimated to be most accurate in part 3, determine its performance on the test data. Is this similar to the cross-validation accuracy estimate? Compare this performance to that of the 1-NN and 3-NN classifiers tested in part 1.

5. Suppose that, instead of using 200 examples per category, we had used 1,000 examples per category. By what factor would the computational cost of classifying test images increase? By what factor would the computational cost of the 5-fold cross-validation procedure in part 3 increase?

2. Linear Algebra

1. Suppose the matrix $A \in \mathbb{R}^{n \times n}$ has n linearly independent eigenvectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$. We denote by M the matrix having these vectors as columns; i.e. $M = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$. Let Λ be a diagonal matrix with the eigenvalues λ_i of $\mathbf{x}_i$ in the diagonal. Show that
$$A = M\ \Lambda\ M^{-1}.$$

2. Compute the eigenvalues and eigenvectors of the following matrix by hand and using Matlab
$$A = \begin{pmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{pmatrix}.$$

3. Let $Q \in \mathbb{R}^{n \times n}$ be an orthogonal matrix, that is it verifies
$$Q^T = Q^{-1}.$$

For any column vector $\mathbf{x} = [x_1 \; x_2 \; \cdots \; x_n]^T \in \mathbb{R}^n$, its Euclidean norm is denoted $\|\mathbf{x}\|$ and is given by

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

Prove that

$$\|Q\mathbf{x}\| = \|\mathbf{x}\|.$$