M532 Mathematical Modeling of Large Data Sets
Problem Set Three

Due Thursday, April 10, 2008

1 Theory

First read Section 4.5.1 in the textbook.

1. Problem 4.2
2. Problem 4.10
3. Problem 4.11 (a) and (b).
4. Problem 4.16
5. Problem 4.17

2 Computing

Select either problem 1. or Problem 2.

Problem 1.

a) (Gappy data warm-up problem). Write a code to i) repair a gappy pattern given a good basis and ii) to construct a good basis from gappy data. Use this code to solve Problem 4.21. How high can you increase the percentage of missing data and still achieve good results?

b) Write a program to repair an image of a cat with 20% of the entries randomly deleted. Apply this repair algorithm both to a cat used to generate the basis in Problem 1. as well as to a cat that was not used to generate that basis. Again, how high can you increase the percentage of missing data and still achieve good results?

c) Extend the code from Problem 2 a) to repair an ensemble of gappy cats. Assume that all the data is gappy and that the basis must be generated iteratively.

Problem 2.

a) (Fisher Fiscriminant Analysis warm-up problem).

- Generate a set of 200 random points in the plane such that half the points lie in quadrants II, III (class 1) and the other half in I and IV (class 2).
- Build a classifier using 80 points from each class using Fisher discriminant analysis.
- Use your classifier to determine which class the 40 points of testing data belong to.

To implement your classifier write a subroutine called "myclassifier" that returns the class labels (variable name testlabels) of the test data (TEST) and the classification error (variable name trainerrors) on the training data (TRAIN). In matlab your call should look like

```
[testlabels, trainerrors] = myclassify(TEST, TRAIN, trainlabels);
```

The variable trainlabels is a vector of 1s and 0s consisting of the class labels of the training data.

b) Repeat the above experiment for the classification data set on the class website. This data sets has 200 columns vectors. The first 100 belong to class 1, the second 100 belong to class 2. Summarize your results in a *confusion matrix* where the rows of the matrix correspond to the actual class of the pattern and the columns correspond to the predicted class.

c) This problem deals with modifying the derivation of linear discriminant analysis done in class.

- If you ignore the within class scatter what is the new optimization problem for linear discriminant analysis, i.e., only include the between class scatter.
- Solve this optimization problem and compare to the eigenvector problem that we derived to generate a best basis.
- Repeat part (b) with this "new" classifier. Note that you can simply write a new subroutine myclassifier2. Compare the confusion matrices and comment on your results.