M532 Mathematical Modeling of Large Data Sets
Problem Set One

Due Thursday, February 21, 2008

1 Theory

1. Text problem 2.15
2. Text problem 2.16
3. Text problem 2.19
4. Text problem 3.8
5. Text problem 3.10
6. Text problem 3.11
7. Text problem 3.21
8. Text problem 3.24
9. Bonus Problem: Text problem 1.5 (See page 346 for definitions of injective, surjective and bijective).

2 Computing

Computing Problem 1: Text problem 2.32.

Computing Problem 2: Text problem 2.33.

Computing Problem 3:

Load the data set on the class web-site into matlab by typing

```
load datamatrix
```

To determine the size of this matrix enter

```
size(datamatrix)
```

Each column is a 64x64 pattern although it is stored as a vector of length 64^2. To unvec and look at the first pattern enter

```
P1 = reshape(Y(:,1),64,64);
imagesc(P1)
colormap(gray)
```

The first 99 columns of Y belong to class 1 (images of cats) and the second 99 columns belong to class 2 (images of dogs).

1. Compute a best basis for the cat images and a best basis for the dog images. Only use 90 images from each set to build these basis.
2. Is there enough data in each set to build a robust basis or would you expect it to change given additional images?
3. Now project the withheld cat and dog images onto the cat and dog bases and by examining the norms of the residuals evaluate the novelty in each case.
4. Compute Shannon's entropy for each data set.
5. Compute the KL stretching dimension with $\delta = 0.01, .1$ and the KL energy dimension with $\gamma = 95\%, 99\%$.