M676 Pattern Analysis
Theory, Algorithms and Computing
Final Project

Thursday, May 10, 2006

This assignment will be collected during our special class on Thursday, May 10 at 9:10pm (Weber 223). Students may work in groups of two (or one) and each group will present their results on May 10. Please select either Project I or Project II as described below.

Note that our last two class sessions on May 2 and May 4 will be in the computer lab Weber 205. I will answer coding questions and give project assistance. Class attendance is expected.

Project I: Prediction

This problem involves the application of the radial basis function approximation technique to the prediction problem. The sunspot cycle data and exchange rate data will be provided to test your algorithm but you may select your own favorite time series as well. Include in your write-up a verification of your algorithm by solving Problem 7.11 in the text.

Employ a time lag of the data such that your RBF model is a mapping $\tilde{f} : U \in \mathbb{R}^3 \to \mathbb{R}$, e.g., approximate the function

$$x_{n+1} = f(x_n, x_{n-1}, x_{n-2})$$

where x_n is the value of the provided time series at time n. Here we have illustrated a lag-3. Select cluster centers (e.g., let $N_c = 10$) using both of the following:

- random subset selection of the data
- LBG clustering of the data (follow algorithm in book)

Use the SVD to solve the over-determined least squares problem. Test your results for each set of centers by attempting to predict future values using both

1. One-step prediction of x_{n+1} where all actual values $x_n, x_{n-1}, \ldots$ known exactly, i.e.,

$$\tilde{x}_{n+1} = f(x_n, x_{n-1}, x_{n-2})$$

2. Iterated prediction of x_{n+1} where the output is used to predict future values via

$$\tilde{x}_{n+1} = f(\tilde{x}_n, \tilde{x}_{n-1}, \tilde{x}_{n-2})$$

Report the means-square error in each case. Explore how your results change as you change the number of centers and time-lags.

Additionally, explore one of the following:

- Explore the behavior of the RBF where 2 centers are chosen to be close together. Illustrate the problems by plotting the singular values of the interpolation matrix. Show how to address this by using the SVD to compute the pseudoinverse of the interpolation matrix.

- Implement the orthogonal least squares algorithm (OLS) as described in class to optimize the number of centers in your model.

Project II: Classification

The goal of this project is to compare a second method of classification to the CCA classification project you carried out in Problem Set 3. Again, your job is to use this data to build a pattern recognition system from one of (or a combination of) the various methods presented in class over the semester.

Submit a brief write-up of each pattern recognition system and include the classification errors as a 2×2 *confusion matrix* matrix (dogs classified as dogs, dogs classified as cats, cats classified as cats and cats classified as dogs). Split your data set into a training and testing data set and provide errors (confusion matrices) for both.

Here are some examples of how you might proceed:

- Fisher's linear discriminant analysis.
- Signal fraction analysis (maximum noise fraction).
- Classification using symmetric eigen-cats and eigen-dogs.
- Wavelet analysis (probably best applied to raw data).
- Fourier analysis, data frequency content.
- k-nearest neighbors (using competitive learning or SOM).
- Radial basis functions (map cats to ones and dogs to zeros).

Elements of your write-up should include the following:

- A simple classification problem to test your method.
- Description of the method used including details of how the classifier is constructed.
- Confusion matrix errors for test and training data.
- Comparison to your CCA classifier.
- Print-out of all code used in project.