Fall 2008 Projects in Applied Math
Radial Basis Function Lab

Instructions: You may work in groups but each student is expected to write his or her own unique Matlab code and complete assignment. The due date is Friday, November 21.

This lab requires you to implement a Radial Basis Function algorithm for predicting time-series data.

Problem 1. Load and plot the time series data. It consists of 2500 values.

Problem 2. Write a subroutine to create time lagged vectors of the data using lags two and four, i.e., vectors of the form

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

and

$$z_n = (x_n, x_{n-1}, x_{n-2}, x_{n-3}, x_{n-4})$$

Problem 3. In this problem, the data set consists of the first 200 points of lag two data. Approximate the function

$$x_{n+1} = f(z_n)$$

using k centers randomly selected from the data plot the error. Take the number of centers to be $k = 5, 10, 15, 20, 25, \dots 200$ and plot the error (sum of the squares of the residuals) of the fit as a function of the number of centers. (Use either $\phi(r) = r$ or $\phi(r) = exp(-r^2)$ as your radial basis functions.) Compute the pseudo-inverse of the interpolation matrix using the SVD.

Problem 4. Repeat this experiment using lag four data. Note, if your computers can handle larger data sets, i.e., more than 200 points, repeat the above with 500 or 1000 points.

Problem 5. Implement the LBG clustering algorithm and repeat Problem 3.

(a) Compute the distortion error as a function of the number of cluster centers.

(b) How do the training errors (on points 1-200) in the model compare with those found in Problem 3?

(c) Compute the error of the model on the *validation points* 201-250 as a function of the number centers. (Only use points 1-200 to build the model.)

(d) An increase in the validation error of the model is a sign of over-fitting. Does your model overfit the data? Discuss.

Problem 6. Test your RBF (LBG clustering) model by computing the normalized prediction error

$$E = \frac{\sum_{n=1}^{P}(x_{n+1} - f(x_n, x_{n-1}, x_{n-2}))^2}{\sum_{n=1}^{P}(x_{n+1} - \overline{x})^2}$$

on points 251-300. (Use the model trained on the first 200 points and validated on points 201-250.) What can you conclude if the error is greater than one?

Problem 7. (*Extra credit.*) Modify your subroutine from Problem 2 to create time lagged vectors of the data using lags two and four, i.e., vectors of the form

$$z_n = (x_n, x_{n-L}, x_{n-2L}).$$

Repeat Problem 6. with the error appropriately modified, i.e., now

$$E = \frac{\sum_{n=1}^{P}(x_{n+L} - f(x_n, x_{n-L}, x_{n-2L})^2}{\sum_{n=1}^{P}(x_{n+L} - \overline{x}))^2}$$