Pattern Analysis Spring 2008
Final Project

Due May 14, 10AM 2008

In this project please employ three to four of the methods below any of the data sets provided.

- Novelty detection
- Singular value decomposition
- Cross validatory SVD
- Missing data algorithm.
- Generalized singular value decomposition
- Signal separation
- Fischer's discriminant analysis
- Canonical correlation analysis
- Radial basis functions
- Topology preserving mappings
- K-means and LBG cluster algorithms
- Discrete Fourier Transform

In particular, focus your efforts on using these methods synergistically.

The data sets include

- Financial Time Series Data
- Kohonen's Animal Data Set
- EEG Data Set (two tasks)
- Cat's and Dog's Data Set
- Mackey Glass time-series
- Exchange rate data

Sample Project I

Compute a radial basis function model for the Mackey Glass equation as

$$x_{n+L} = f(x_n, x_{n-L}, x_{n-2L}, x_{n-3L})$$

where $L = 16$. Train on the first 1000 points and validate your model using points 1001-1200.

a) Test your model by computing the normalized prediction error

$$E = \frac{\sum_{n=1}^{P}(x_{n+L} - f(x_n, x_{n-L}, x_{n-2L}, x_{n-3L}))^2}{\sum_{n=1}^{P}(x_{n+L} - \overline{x})^2}$$

on points 1201-1300.

b) Explore the model by varying the number of centers. Propose a method to automatically determine the order of the model.

c) Test this approach on the exchange rate data but in addition to predicting using a raw time series use the first basis vector from signal fraction analysis as the time series. Do your prediction errors improve?

Sample Project II

This project concerns assessing the quality of the missing data algorithm as measured by alterations in canonical correlations and clustering distortion error. Evaluate the missing data algorithm using the following steps:

a) Delete 10%, 20% and 50% of the pixels at random in the cats and dogs data set. Repair these data sets using the missing data algorithm.

b) Compute the canonical correlations between the corrupted cats and dog data sets, as well as the repaired cats and dogs data sets and compare with the original CCA analysis.

c) Employ an LBG clustering algorithm on all seven cat/dog data sets (i) original (ii) corrupted and (iii) repaired. Plot the final distortion error in each case using 1, 2, ..., 50 clusters. We expect that using 50 clusters on the dog or cat data set should produce relatively low distortion errors. Is the distortion error lower on the repaired data sets than on the corrupted data sets?