M532 **Spring 2018**
Problem Set 6. Due Sunday, May 13 at 12 noon

Part A: Theory and Foundations

Problem 1

The RBF linear system requires the solution of the overdetermined least squares problem

$$y = \Phi w$$

a) Show that by applying the operator Φ^T to both sides of the equation that

$$w = (\Phi^T \Phi)^{-1} \Phi^T y$$

Note that this derivation is different from the one provided in class, don't reproduce that here.

b) Verify that the actual output of the model is $\overline{y} = \Phi w$ where

$$\bar{y} = \Phi(\Phi^T \Phi)^{-1} \Phi^T y$$

c) Let $\mathbb{P} = \Phi(\Phi^T \Phi)^{-1} \Phi^T$. Show that $\mathbb{P}$ is an orthogonal projection, i.e.,

i) $\mathbb{P} = \mathbb{P}^2$ (projection criterion)

ii) $\mathbb{P} = \mathbb{P}^T$ (criterion for projection to be orthogtonal)

Problem 2

Show how the problem

$$y = \Phi w$$

can be solved using

a) the SVD

b) the QR decomposition

Problem 3

Consider the set of input-output pairs $\{(0,1),(2,4),(3,8),(4,12),(5,17)\}$. Let the model for this data be

$$f(x) = a\sqrt{x} + bx^{3/2}$$

Find a and b.

Part B: Computing

Problem 1

a) Load and plot the time series data from canvas. This temperature data was collected by the CSU Weather Station over September 2017 at 5 minute intervals.

b) Write a subroutine to create time lagged vectors of the data using time-delay embedding, i.e., vectors of the form

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

and plot the resulting data in three dimensions. Experiment with different values of T.

In the RBF and ANN problems you will divide your data into training, validation and testing sets. The training set is used to build the model. The error on the validation set is used to determine the number of centers. The error on the test set is a measure of the quality of the model.

Problem 2

Model this time series using RBFs. The spirit of this problem is to explore various options, e.g., which RBF should you choose? How many centers? How large should the training, and validation sets be? Use the following approaches:

- Select your centers randomly from the data.
- Select your centers using LBQ clustering.

Evaluate the quality of your model as a function of the number of centers in each case using the error

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

Include plots of the error on the validation and training sets as a function of the number of centers. Save the last 500 points of your data for testing the model with this error. After you have determined the number of centers in your RBF using the training procedure above, compute the error on your test set.

Problem 3

Repeat problem 2 using the multilayer perceptron. Use both the hyperbolic tangent and the relu function as the nonlinear transfer function and compare. You will need to determine the number of nodes in the hidden layer by considering the error on the validation set. Include a plot of the training error and validation error as a function of the number of nodes in the hidden layer of the network. Report your error on the test set after you have determined the number of nodes using your validation set.

Problem 4

Based on your results in Problems 2 and 3 which modeling approach would you recommend and why?

Problem 5

a) Using the Indian Pines Hyperspectral Data Set obtain 25 points on the Grassmannian $Gr(k, n)$. These points should not share any samples.

b) For each class compute the class mean of your subspaces.

c) Build a distance matrix using the chordal metric applied to the data and the class means.

d) Use your MDS code to visualize the Grassmannian in two dimensions. Clearly label each class and the class means.

Repeat the above taking $k = 1, 5, 10$.

Problem 6

Extra Credit Implement the OLS algorithm for RBFs and compare the resulting model with what you found in Problem 2.