**Exam 2 Notes**
Math 350
Spring 2008

You can prepare 1 page (single sided) of notes to bring to the exam with you. You should also bring a calculator. The second exam covers the Best Basis, Data Clustering (LBG, Neural Gas, SOM), polynomial interpolation and regression (including linear networks), RBFs, and some nonlinear networks (Tuesday's class). I won't test you over the non-orthogonal basis material (Chapter 7).

The exam will be 70% in class, and 30% take home (Matlab). You are expected to do your own work for the take-home exam (it will be due on Tuesday April 29th at 5PM).

- Definitions: Best basis, SVD, linear neural network, lag, LBG, SOM, RBF, Voronoi cell, Neural Network, OLS, Covariance matrix (from a matrix $X$), distortion error, transfer function, interpolation, regression, EDM, Vandermonde matrix.
- Know the learning rules for the LBG, SOM and the Neural Gas.
- Linear Algebra techniques we use:
  - Project $\mathbf{x}$ onto $\mathbf{y}$. Remove the component in the direction of $\mathbf{x}$ from the vector $\mathbf{y}$ (this last technique is used in Orthogonal Least Squares).
  - Find the eigenvalues and eigenvectors of a symmetric matrix.
  - The relationship between the SVD and the best basis for the column space and the row space of a matrix.
  - What metric is the "Best Basis" algorithm using to get a *best* basis?

Below are some sample problems to consider:

1. True or False, and give a short reason:
   (a) We talked about two ways of choosing centers for the RBF. We can either choose them at random or use OLS.
   (b) The EDM is invertible.
   (c) (Review question) Given a column vector $\mathbf{x}$ in $\mathbb{R}^n$, the matrix:
   $$\frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{x}}$$
   is a projector, and it is an orthogonal projector.
   (d) The LBG algorithm finds a (local) minimum to a known error function.

(e) If $\mathbf{x}$ and $\mathbf{v}$ are column vectors in $\mathbb{R}^n$, then to remove the components of the columns of $Y$ in the direction of $\mathbf{x}$ and of $\mathbf{v}$, we compute:

$$\left( I - \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{x}} - \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}} \right) Y$$

2. Short Answer:

   (a) Define a Voronoi Cell:

   (b) Suppose we wish to visualize a high dimensional data set on a computer monitor. Then Neural Gas is probably the most reasonable way to cluster the data.

   (c) What is the difference between interpolation and regression?

   (d) Given a $p \times n$ matrix $X$, give the algorithm (as a series of steps- Use English or Matlab) for how we can determine the best basis for the columnspace of $X$. Does the algorithm change for the rowspace?

   (e) What function are we attempting to find when we perform data clustering?

   (f) Is data clustering supervised or unsupervised learning?

   (g) What is the curse of dimensionality?

   (h) Converting a time series (a big row vector) into a matrix with $k + 1$ rows is called what? Hint: For example,

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$$

   (i) How many parameters are there in a degree two polynomial in four variables (e.g., $r = f(x, y, z, w)$)?

   (j) Before we can train the RBF, there are basically two questions about the model that must be answered first. What are they?

   (k) In the RBF model, we can always improve the accuracy of the model by adding more centers. Why don't we always add enough centers to get zero error?

   (l) When finding the best basis, we may need to decide how many basis vectors we should keep. How might that be done?

3. Show, using polar coordinates, that:

$$\left( \int_{-b}^{b} \mathrm{e}^{-x^2}\, dx \right)^2 \approx \int\!\!\int_B \mathrm{e}^{-(x^2+y^2)}\, dB = \pi(1 - \mathrm{e}^{-b^2})$$

   where $B$ is a disk of radius $b$.

4. (a) If $C$ is the covariance matrix given below, find the maximum of $F(\phi)$, and give the $\phi$ for which the maximum occurs (we may assume $\phi$ is not the zero vector, and that $\phi$ is a vector with 2 elements).

$$F(\phi) = \frac{\phi^T C \phi}{\phi^T \phi} \qquad \text{for } C = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

(Hint: You may find it easily using our theorems)

(b) Find the minimum of $F$.

5. Here are four data points in the matrix $X$, with the corresponding class labels in $Y$. We will use a linear neural network to perform the classification.

$$X = \begin{bmatrix} 2 & 1 & -2 & -1 \\ 2 & -2 & 2 & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} -1 & 1 & -1 & 1 \end{bmatrix}$$

If you initialize the weights to ones, and the bias(es) to zero, perform two updates of the weights and biases according to the Widrow-Hoff algorithm using data points chosen at random (that would be the third data point, then the second data point).

6. Using the previous question, write the system of equations you would use to solve the classification problem. Will the solution you find actually be a solution? Explain.

7. Here are five data points in the matrix $X$. Using the LBG algorithm, initialize our set of two centers using data in columns 1 and 2, and perform one update:

$$X = \begin{bmatrix} -1 & 1 & 1 & -2 & -1 \\ 1 & 0 & 2 & 1 & -1 \end{bmatrix}$$

Show that there is a decrease in the distortion error.

8. Here is one data point. There are three centers in the matrix $C$ which have a linear topology- That is, $I$ gives the one-dimensional representation of each cluster center.

Perform one update of the centers using Kohonen's SOM update rule, assuming that $\epsilon = \lambda = 1$ (unrealistic, but easier to do by hand):

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad C = \begin{bmatrix} -1 & 1 & 2 \\ 1 & 0 & 3 \end{bmatrix} \qquad I = [1, 6, 3]$$

9. Same as the previous problem, but update using the Neural Gas algorithm (assume all the centers are connected and ignore the age). Use $\epsilon = \lambda = 1$ (unrealistic, but this is by hand).

10. Given that $\{v_1, v_2, \ldots, v_k\}$ are the (orthonormal) vectors that form the best basis for some fixed set of $p$ points $\mathbf{x}_1, \ldots, \mathbf{x}_p$ in $R^n$, is there a way of indirectly computing the mean square error in using the $k$ basis vectors? (Hint: eigenvalues)

11. Here is some data:

| $x$ | $-1$ | 0 | 1 | 2 |
|---|---|---|---|---|
| $y$ | 1 | 0 | 1 | 3 |

Write the matrix-vector system we would solve in Matlab if we were modeling this using:

- A line.
- A parabola.
- An interpolating polynomial.
- An RBF using the first and third points of $x$ as centers, and a cubic transfer function.

12. In the SOM, once the clustering algorithm has finished, we have two mappings. What are they, and give an example of when you would use each. (Hint: Think of how we used the SOM in the homework).

13. Let $A$ be $m \times n$ with rank $k$. If $A = U\Sigma V^T$ is the thin (or economy) SVD, what is the pseudoinverse of $A$ (in terms of $U, V, \Sigma$)? Be explicit about the sizes of the matrices involved.

    *Side Note:* This is what Matlab computes when we called `pinv(A)` in our RBF routines.

14. In Problem 8, which column of $C$ best points in the direction of $\mathbf{x}$?

15. Given a fixed set of $p$ points, $\mathbf{x}_1, \ldots, \mathbf{x}_p$, show that the vector $\mu$ that minimizes:

$$E(\mu) = \sum_{j=1}^{p} \|\mathbf{x}_j - \mu\|^2$$

    is the mean.