Homework From Week 9: Linear Nets, Stats

1. **Breast Cancer Data Classifier**

   Use the Iris Data as a template, and build a linear network that will classify the data given on the class website as `BreastData.mat`, which is described below.

   Each data "point" represents nine measurements taken from a breast exam, and there are six target classes:

   - Class 1: Carcinoma
   - Class 2: Fibro-adenoma
   - Class 3: Mastopathy
   - Class 4: Glandular
   - Class 5: Connective
   - Class 6: Adipose

   If you would like more information about this data, see "Variability of impedivity in normal and pathological breast tissue", by J. Jossinet. Med. & Biol. Eng. & Comput, 34: 346-350 (1996).

   If you download the Matlab $m-$file (this is NOT a binary data file, but a plain text script), you can see a few notes there. To load the data into Matlab, just type `BreastData` in the command window (or in a script) to load the matrices X and T. Notice that they may not be given to you with the right dimensions (the data may be in columns or in rows), so be sure to check that and use the dimensions appropriate for your algorithm.

   Using this data, construct two script files- One using the Widrow-Hoff update (you can use the `WidHoff.m` function on the class website), and one that finds the weights and biases using batch training. Both scripts should "output" the final confusion matrix (just leave the semicolon off of the last computation).

   Think about what the numbers mean in terms of the classes we've described, and give a short summary.

   From the Statistics chapter, Exercises 10 and 11, written again below.

   For both sets, suppose $\mathbf{x}$ is a vector containing n real numbers, and we understand that $m\mathbf{x} + b$ is Matlab-style notation (so we can add a vector to a scalar, done component-wise). Use the results of exercises 6-9 to help you.

2. Show that, for fixed constants a, b, $\text{Cov}(\mathbf{x} + a, \mathbf{y} + b) = \text{Cov}(\mathbf{x}, \mathbf{y})$

3. As a summary of question 11, if $\mathbf{y} = m\mathbf{x} + b$, then find the covariance and correlation coefficient between $\mathbf{x}$ and $\mathbf{y}$.

(From the Linear Regression Section, last page)

4. Find the line of best fit through the data found in the table on p. 39 of the linear regression notes (this is the Hanford data relating an "index" to the number of deaths). Show the result by graphing the data and the line you found.

   Since A^TA is invertible, you can solve the equation given, $A\mathbf{c} = \mathbf{t}$ using Matlab. The plotting commands are also given below, assuming the slope is the first coordinate of $\mathbf{c}$ and the intercept is the second coordinate, and the "index" is the first column of A.

```
x=...  %Enter the data as a column vector
t=...  %Enter the data as a column vector
A=...  %Construct the matrix A
c=inv(A'*A)*A'*t;
xx=linspace(min(x),max(x));
yy=c(1)*xx+c(2);
plot(x,t,'r*',xx,yy,'k-');
```

5. Be sure you understand how to apply the Widrow-Hoff update rule. For example, suppose W is the 2×2 identity matrix, and the other vectors are given by the following, and the learning rate $\alpha = \frac{1}{10}$.

$$\mathbf{b} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \mathbf{t} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

   Perform one step of the Widrow-Hoff rule to update W and $\mathbf{b}$.