CAP 6610 Machine Learning
Homework 2 (due on October 16):

1. (Backpropagation)
   Consider the following network: 1 input layer, 1 hidden layer and 1 output layer, each layer with 2 neurons.

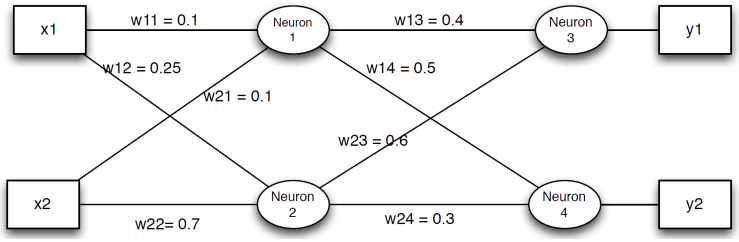

Figure 1:

   You can assume you are using L2 for loss function: $J(w) = \frac{1}{2}\sum_{i=1}^{N} e_i^2$ where e=d-y
   All weights are initialized to the values shown (and there are no biases for simplicity). Consider the data point $x = [1, 1]^T$ with desired output vector $d = [1, 0]^T$. Complete one iteration of backpropagation by hand assuming a learning rate of $\eta = 0.1$.
   Use the sigmoid activation function: $\phi(x) = \frac{1}{1+e^{-x}}$. Note that $\frac{d\phi(x)}{dx} = \phi'(x) = \phi(x)(1 - \phi(x))$.
   Question: What would all the weight values between input layer and hider layer be after one backpropagation iteration? (w11, w12, w21, w22)

2. (Backpropagation)
   Still according to the same condition from question 1.
   Question: What would all the weight values between hider layer and output layer be after one backpropagation iteration? (w13, w14, w23, w24)

3. As following figure, suppose you want to create a 2-hidden layer network to distinguish the letters (FL) and the background.

Figure 2:

How many units will you need in the first and second hidden layers? Why? Justify your answer by providing an explanation of each hidden unit role in creating this network. (Just write your answer. Not necessary to code a real model)
Can you achieve the same goal with a single hidden layer network? Why or why not?

4. Suppose you would like to learn the number of neurons you need in a neural network with a single hidden layer by beginning with too many neurons and, then, driving the weights for unnecessary neurons to zero through the use of an appropriate regularization term. Considering the following two regularization terms,

$$E_1 = \sum_{i=1}^{M} w_i^2$$
$$E_2 = \sum_{i=1}^{M} |w_i|$$

which one would be better? Why?

5. In a presumably unfair dice the unknown probabilities of appearance of the individual faces are respectively $p_1, ..., p_6$ (none of them are zeros). Through repeated experiments it has been noted that the probability of rolling a streak of 6 consecutive numbers (i.e. rolling a sequence 1,2,..., 6) is greater than or equal to: $(\Pi_i p_i^{p_i})^{n/\Sigma_i p_i}$. What are the possible values of $p_1, ..., p_6$?

6. Alex has built a Multi-layer Perceptron which consists of a single hidden layer of neurons. The model evaluates a boolean function f on four input variables X_1, X_2, X_3, X_4. Also it is known that each neuron in the network is simple enough and acts as a boolean gate. Alex observed that working with this simple network, he ends up using maximum number of neurons that is necessary to evaluate any boolean function in four variables. What is the number of neurons that Alex uses? If Alex is allowed to use any number of layers (i.e. he could increase the depth of the network) what would be minimum number of neurons required to model the same function? Can you say something about the importance of depth in an MLP from this example?

7. Code and test a two layer feed-forward net of sigmoidal nodes with two input units, ten hidden units and one output unit that learns the concept of a circle in $2D$ space. The concept is: $\langle x, y \rangle$ is labeled $+$ if $(x-a)^2+(y-b)^2 < r^2$ and is labeled $-$ otherwise. Draw all data from the unit square $[0,1]^2$. Set $a = 0.5, b = 0.6, r = 0.4$. Generate 100 random samples uniformly distributed on $[0,1]^2$ to train the network using error back-propagation and 100 random samples to test it. Repeat the procedure multiple epochs and with multiple initial weights. Report the changing accuracy and the hyperplanes corresponding to the hidden nodes (when the sigmoid is turned into a step function).