YOUR NAME:
YOUR ID NUMBER:

MIDTERM FOR CSC 321

Time: Tuesday Feb 24, 1.10-2.00pm
CLOSED BOOK TEST

Answer ALL FIVE questions in Part A. Answer exactly 2 questions in part B.
Each question in part A is worth 2 points. Each question in part B is worth 5 points.

PART A

1. Describe one way to speed up learning when the horizontal cross-sections of the error surface are elongated ellipses. Explain why the method works.

2. Fill in the three weights on the hidden-to-output connections and the biases of the hidden and output units so that this network gives an output of 1 if and only if the binary input vector has odd parity *(i.e. an input of 101 should produce an output of 0)*. Assume that the hidden and output units are binary threshold units: They give an output of 1 if their bias plus the total input they receive from below exceeds zero and an output of zero otherwise.

3. a) The Bayesian equation for the posterior probability of a parameter, θ, is:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} \tag{1}$$

Derive this equation from the equation that relates a joint probability to a conditional probability.

b) Explain very briefly the difference between Maximum Likelihood learning (ML) and Maximum a Posteriori Learning (MAP).

4. Explain precisely why early stopping prevents overfitting.

5. a) Briefly describe one way in which a “Mixture of Experts” is better than just averaging the results of many different expert networks that are all trained separately.

b) Briefly describe one way in which a “Mixture of Experts” is worse than just averaging the results of many different expert networks that are all trained separately. Assume that the different experts that are averaged together each have the same structure as the individual experts in the mixture of experts.

PART B: Answer EXACTLY 2 questions in this part.

(*Make sure to cross out answers that you do not want to be graded. If you do not, your worst two answers will be graded.*)

1. a) *(3 points)* There are 4 balls in an urn. Each ball is either red or black. You start by believing that the probabilities that the urn contains 0, 1, 2, 3 or 4 red balls are all equal. You then reach into the urn and pull out one ball at random. It is red. You put this ball back, then you reach in again and pull out a ball at random. It is black. Show how your probability distribution for the number of red balls changes when you pull out a red ball and show how it changes again when you pull out a black ball.

   b) *(2 points)* In full Bayesian learning, how do we use the posterior distribution over parameters after we have computed it?

2. *(3 points)* Explain exactly what is wrong with the following "proof" of the perceptron convergence theorem :

   "Each time we select a training case that gives the wrong answer, we increment the weight vector by the input vector if the desired answer is 1 and we decrement the weight vector by the input vector if the desired answer is 0. The change in the weight-vector always reduces the distance between the current weight vector and any weight vector that produces the right answer for all training cases."

   (You can Assume that we cycle through the training set selecting cases in a fixed order. You can also assume that all the input vectors are binary.)

   b) *(2 points)* Suppose we take a perceptron that is learning a training set that it can get perfectly correct. Suppose it has not yet learned to get all of the training cases right. If we multiply all of the weights by 2 and we also multiply the bias by 2, does this affect the error rate of the perceptron on the training set?

3. a) (*2 points)* How does self-supervised backpropagation in a linear network differ from principal components analysis?

b) (*3 points)* How can self-supervised backpropagation be used for clustering data?

4. a) (*2 points)* How do weight contraints enable LeNet to get better generalization performance than networks that do not use weight constraints.

b) (*2 points)* Why does LeNet have several different pools of replicated hidden units in the first hidden layer of the network?

c) (*1 points)* The information that LeNet captures by using weight constraints can also be captured in a brute force way without using weight constraints. What is this brute force method?