STUDENT NAME: _____

STUDENT ID: _____

MIDTERM EXAMINATION

Machine Learning - Fall 2006

November 1, 2006

You are allowed one double-sided "cheat sheet"

Read all the questions before you start working. Please write your answer on the provided exam. Partial credit will be given for incomplete or partially correct answers. Please be sure to define any new notation you introduce.

Good Luck!

1. [25 points]

   For each of the learning problems outlined below, specify what is the best learning algorithm to use and why. Note that you should give *one* algorithm for each problem, even if there are several correct answers.

   (a) You have about 1000 training examples in a 6-dimensional continuous feature space. You only expect to be asked to classify 100 test examples.

   (b) You are going to develop a classifier to recommend which children should be assigned to special education classes in kindergaren. The classifier has to be justified to the board of education before it is implemented.

   (c) You are working for a huge retailing company. You are trying to predict whether customer X will like a particular item, as a function of the input which is a vector of 1 million bits specifying whether each of the other customers liked the item. You will train a classifier on a very large data set of items, where the inputs are everyone elses preferences for that item, and the output is customer Xs preference for that item. The classifier will have to be updated frequently and efficiently as new data comes in.

   (d) You are working for a geological research institute. You get data consisting of 10 continuous measurements of properties of a rock sample, and an amount of iron present in the rock. you need to use machine learning to be able to predict the amount of iron in newly received rocks.

   (e) You are working for an oil company which is trying to decide where to drill. You have 40 attributes, both discrete and continuous, that describe a plot of land. Some of these attributes

are noisy. For previous sites, you know whether they contained oil or not, but you only have data about 50 such sites.

2. [15 points] **Short questions**

   (a) For a neural network, which of the following affects the trade-off between underfitting and overfitting (circle all that apply)

      i. the number of hidden units

      ii. whether simple backpropagation or second-order methods are used to update the weights

      iii. the number of hidden layers

      iv. whether the data you have has noisy outputs

      v. whether the data you have has noisy inputs

   (b) True or false: a large enough neural network with access to enough data will learn any function.

   (c) True or false: If a set of hypothesis can shatter a set of instances D_1 of size k but cannot shatter a set of instances D_2 of size $k + 1$, then the VC dimension is k.

   (d) True or false: a support vector machine

   (e) You have been asked to review a paper on machine learning applied to dragonology. The authors had 20 data sets, and tested their algorithm against three others. They report the error on the training set and it is always the smallest. Will you accept the paper? Explain in 1 sentence why or why not.

3. [15 points] You are given a set of points in 2D, with the following coordinates and labels:

$$(1, 1, -)(1, 2, -)(1, 4, +)(3, 1, +)(3, 2, +)$$

   (a) Draw a picture of all the possible instances. Draw a decision tree that separates the data. Draw the boundary between instances provided by the tree.

   (b) Does an optimal separating hyperplane exist? If so, draw it and give its equation. If not, explain why not.

   (c) Draw the decision boundary corresponding to the 1-nearest neighbor algorithm

4. [10 points] **Gradient descent**

   Suppose that you are training a regression approximator of the form:

$$h(x_1, x_2) = w_0 + w_1 \sin(w_2 x_1) + w_3 \sin(w_4 x_2)$$

   Derive the gradient descent update rule for all the parameters.

5. [10 points] **VC dimension**

   Suppose that you have data in 2D. What is the VC dimension of rectangles with axis-parallel edges? the examples inside the rectangle should be classified as positive. Justify your answer.

6. [10 points] **Kernels**

   Let $K(x, y)$ be a kernel function and $a, b \in \Re$. Under which conditions is the function $aK(x, y) + b$ a kernel?

7. [15 points] **Learning functions**

   Consider learning random boolean functions under the following setup.

   - Target functions with k binary inputs and 1 binary output are generated by randomly assigning outputs (independently with probability one half) to each possible combination of inputs.
   - Noiseless training data is generated by randomly selecting a setting of the inputs (uniformly), reporting its output, and repeating this process N times independently.
   - Test data inputs are generated by randomly selecting a setting of the inputs (uniformly) and repeating this process M times independently.

   (a) Let a be the expected number of distinct training cases in a training set [Obviously $1 \leq a \leq \min(N, 2k)$]. Derive an expression for a in terms of k and N. What is the expected number b of cases in the test set which also appeared in the training set? (In terms of a, k, M.)

   (b) What is the lowest possible expected test set error rate we can hope to achieve? (In terms of b, M.)

   (c) Give an example of an algorithm which achieves this error rate and an argument (in ¡25 words) of why no algorithm can do better.