STUDENT NAME: ______________________________

STUDENT ID: ______________________________

MIDTERM EXAMINATION

Machine Learning - Fall 2004

November 4, 2004

This examination is open book, open notes.

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

Good Luck!

1. [40 points] A foreign government has recently hired you as a consultant in order to predict the results of absentee ballots from Canada. There are two candidates for presidency. A pool of 1000 people have agreed to complete a post-vote survey, recording their age, education level (no degree, high-school, certification, college or graduate-level), faith (one of 20 categories), marriage status, how many children they have and the household income level (less than 20k, between 20k and 50k, between 50k and 100k or over 100k), and what vote they cast. Most people will be truthful, but some will not. The personal information above is available for all voters residing in Canada, but some values might be missing or outdated. You looked at your old lecture notes and made a list with most of the algorithms that were discussed:

   (a) Decision trees

   (b) Linear regression

   (c) Polynomial regression

   (d) Logistic regression

   (e) Feed-forward neural networks

   (f) 1-Nearest neighbor

   (g) 5-Nearest neighbor

   (h) Weighted nearest neighbor

   (i) Support vector machines

   (j) Naive Bayes

   For each of these methods, state whether it is applicable to this problem (yes or no). If you answer no, explain why not. If you answer yes, explain why, and describe any necessary implementation details (e.g. how the classifier would be structured, choices of input encoding if necessary, distance functions, parameters etc). Do NOT write more that 3-4 sentences for each method!

.

2. [2 points] Bayes rule is a popular tool in machine learning. Your friend claims to know that Google searches the web using a neural network which makes heavy use of Bayes rule. Why are you suspicious? Circle the answer that applies (no explanation is necessary)

   a) Bayes rule cannot be applied when dealing with large amounts of data (such as pages on the web)

   b) Neural networks are usually based on information gain instead of Bayes rule

   c) A nearest neighbor approach would be much more effective than a neural network with Bayes rule

   d) Bayes rule is usually used in connection with decision trees instead of neural networks

   e) To the best of your knowledge, Bayes rule has nothing to do with neural networks

3. [2 points] You are given a learning task where you must predict a probability value $\in [0,1]$ based on 5 binary attributes. The training set consists of 1,000 instances and the time available for both learning and classification is very limited. Circle the best method to use from below (no explanation is necessary)

   a) neural networks because the training set is large

   b) logistic regression because it is an eager learning method

   c) naive Bayes because this is an eager learning method

   d) radial basis function network because we are predicting a probability

   e) nearest neighbor because it is instance-based and thus requires little time

4. [2 points] You are working on a supervised learning task with 2 continuous attributes and a continuous output that ranges from 0 to 100. Some of the methods that you would like to use include:

   1) neural network
   2) linear regression
   3) polynomial regression
   4) naive Bayes
   5) nearest neighbor

   Which of the above could you try? Circle all the methods that apply (no explanation is necessary)

5. [4 points] Suppose you are trying to learn a function y of one input x, using gradient descent as we have seen many times in class. You have a hypothesis of the following form:

$$y = w_1 + w_2 \log(w_3 x)$$

What are the update rules for w_1, w_2 and w_3?

6. [10 points] **A bit more than naive Bayes**

   Suppose that instead of the naive Bayes model, you have a network in which each attribute also depends on the one preceding it, like in the figure below.

   (a) How many parameters are necessary to describe your model?

   (b) Given a set of instances $\langle \mathbf{x_j}, y_j \rangle, j = 1 \ldots m$, describe how these parameters would be computed based on the values of $x_{i,j}$ and y_j.

7. [10 points] Suppose that you are trying to learn the following function of two Boolean variables:

$$f(A, B) = \neg A \wedge B$$

   (a) Draw a picture of all the possible instances
   (b) Draw the corresponding decision tree
   (c) Draw a perceptron or a network of perceptrons, as appropriate, which separates the training instances. Specify the necessary weights.
   (d) Does an optimal separating hyperplane exist? If so, draw it and give its equation. If not, explain why not.
   (e) Draw the decision boundary corresponding to the 1-nearest neighbor algorithm

8. [10 points] **Decision trees**

   Suppose that you are given a binary classification problem, with binary attributes $A_1, \ldots A_n, B_1, \ldots B_n$ and two classes C_1 and C_2. An expert gives you a rule set of the form:

$$\begin{aligned} A_1 \wedge B_1 &\Rightarrow C_1 \\ A_2 \wedge B_2 &\Rightarrow C_1 \\ &\vdots \\ A_n \wedge B_n &\Rightarrow C_1 \\ T &\Rightarrow C_2 \end{aligned}$$

   In other words, you have a set of rules for classifying examples as class C_1, and if none of them applies, the example will be given class C_2 by default.

   (a) Draw the decision tree that corresponds to this set of rules

   (b) How many internal nodes and how many leaves does the tree have?

9. [10 points] **Neural networks**

A medical insurance company hires you as a consultant to help them classify which patients should receive an expensive treatment and which should not. The company has access to a variety of patient data (personal data, symptoms, medical tests, family history...). If the company chooses to treat a patient, and the patient is not really sick, the cost of the treatment is waisted. In they choose not to pay for someone who is really sick, the patient might die and the company would have to pay a huge amount of money to the family. In other words, not all classification mistakes are equal

(a) Can the cost information be incorporated in the training algorithm for the neural network? If so, describe how. If not, explain why not

(b) List your top 3 concerns regarding the application of neural networks to this problem.

(c) If you would prefer a different learning method, name it and explain in one sentence why it would be preferable.

10. [10 points] **Nearest neighbor** Suppose you have 100 data points (x_k, y_k), generated as follows: $x_k = \frac{k}{100}, k = 1, \ldots 100$, and y_k is drawn from a Gaussian distribution of mean 0 and $\sigma^2 = 4$. All values y_k are drawn i.i.d. Consider two algorithms: 1-nearest neighbor and always-0 prediction. Compute the expected mean squared training error for the two algorithms. Compute the mean squared testing error for leave-one-out cross-validation for the two algorithms.