STUDENT NAME: _____

STUDENT ID: _____

MIDTERM EXAMINATION

Machine Learning - Fall 2005

October 27, 2005

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. [50 points]

   A financial institution has just hired you to build a system which will decide what car insurance package to offer to different clients. The information recorded about the clients is; gender (Boolean), age group (under 20, 20-35, 35-55, over 55), occupation (one of 20 categories), credit rating (a numerical value between 0 and 20) and number of accidents in the last year, and in the last five years (as reported by the client themselves). They also record the car make, model and year, and the city where the person lives. They have a database of roughly 20000 current clients, and for each they have recorded the plan that was used, as suggested by an expert.

   For each of the methods below, explain whether it is appropriate to use for this type of data. If your answer is no, explain why. If your answer is yes, explain why, and how you would go about setting up the problem (input processing, output processing, choice of approximator structure etc).

   (a) Neural networks

   (b) Support vector machines

   (c) Linear regression

   (d) Polynomial regression

   (e) Naive Bayes

(f) Decision trees

(g) k-nearest neighbor

(h) Weighted nearest neighbor

(i) Logistic regression

(j) Bagged decision trees

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

   (a) [3 points] For polynomial regression, which of the following affects most the trade-off between underfitting and overfitting (no explanation necessary):

      i. the degree of the polynomial
      ii. whether the weights are computed by matrix inversion or gradient descent
      iii. the variance of the Gaussian noise
      iv. the use of a constant-term unit input

   (b) [5 points]
   Suppose you are trying to predict a class label Y based on two binary input X_1 and X_2. Draw the naive Bayes classifiers for this problem. How many parameters will have to be learned?

   (c) [2 points[Is the following statement true or false?
   If there exists a set of instances of size k that *cannot* be shattered by a hypothesis class H, then $VC(H) < k$.

   (d) [2 points] True or false: The training error provides a pessimistic estimate of the true error of a classifier (no explanation necessary)

   (e) [3 points] True or false: a linear support vector machine will find the globally optimal hyperplane with respect to its optimization criterion

3. [10 points] Suppose that you are trying to learn the following function of three Boolean variables:

$$f(A, B, C) = A \vee (B \wedge C)$$

   (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

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

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

   $$h(x_1) = w_0 + w_1 x_1 + w_2 x_1^2$$

   Give the gradient descent update rule for w_0, w_1, w_2

   (b) Now suppose the approximator is of the form:

   $$h(x_1) = w_0 + w_1 x_1 + \log(w_2 x_2)$$

   Give the gradient descent update rules only for the weights for which the answer is different from the previous part.

5. [7 points] **VC dimension**

   Let H_1, H_2, H_3 be three hypothesis classes such that $H_1 = H_2 \cup H_3$. Is it true that $VC(H_1) \leq VC(H_2) + VC(H_3)$, or not?. Justify your answer.

6. [8 points] **Kernels**

   Consider the Gaussian kernel:

$$K(\mathbf{x_i}, \mathbf{x_j}) = e^{-\frac{||\mathbf{x_i}-\mathbf{x_j}||^2}{2}}$$

   . As we discussed in class, there is some mapping ϕ such that $K(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i}) \cdot \phi(\mathbf{x_j})$. Prove that, for any two input instances $\mathbf{x_i}$ and $\mathbf{x_j}$,

$$||\phi(\mathbf{x_i}) - \phi(\mathbf{x_j})||^2 < 2$$