STUDENT NAME: _____

STUDENT ID: _____

MIDTERM EXAMINATION

Machine Learning - Fall 2007

October 31, 2007

This is an open-book, open-notes exam. No laptops are allowed.

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. [32 points] What can you apply?

   A candy factory is preparing for Halloween and they want to optimize their production pricing. They currently produce 3 types of candy: Apple, Banana and Chocolate. They have data from 10 previous years regarding their production price, the price at which they sold the candies, and the quantities that were bought. They also have data about the total number of candies of these 3 types that were on the market in the past 2 years, and they know the retail price charged for competitor candies in the last 10 years (but not the quantities that were sold). They want you to help them predict the amount of candy of each category that they will be able to sell, as function of the prices asked. Which of the following machine learning algorithms can you use? For each one, if not applicable, explain in one sentence why. For each one that is applicable, explain in at most 3 sentences why it is applicable, and whether it would require any pre-processing of the data.

   (a) Logistic regression

   (b) Linear regression

(c) Polynomial regression

(d) Weighted linear regression

(e) Neural networks

(f) Support vector regression

(g) k-nearest neighbor

(h) Regression tree

2. [28 points] **Short questions**

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

      i. the regularization constant
      ii. the number of instances in the training data
      iii. the number of attributes in the training data
      iv. the choice of kernel

   (b) Consider below 4 methods used to classify 2D data. Rank them from 1 to 4 where 1 is highest bias and 4 is lowest bias:

      - 1-nearest neighbor classifier
      - Logistic regression
      - Boosted logistic regression with 10 rounds of boosting
      - Decision trees of depth 10

   (c) True or false: A classifier trained on more training data is less likely to overfit

   (d) True or false: In boosting, the weights of misclassified examples go up by the same amount in one round of boosting

   (e) Would a point that is far away from the decision boundary influence the weights of a classifier learned by logistic regression or not? Explain your answer in one sentence.

   (f) Consider an SVM and suppose that we remove one of the support vectors from the training set. Will the margin increase, decrease or stay the same for this data set?

   (g) Consider an SVM and suppose that we remove from the training set a point which is not a support vector. Will the margin increase, decrease or stay the same for this data set?

3. [5 points] **Nearest neighbor and decision trees**

   Suppose you have data with 2 real-valued attributes. Is there any relationship between the decision boundary of a 1-nearest neighbor classifier and that of a decision tree (assuming no pruning)? Justify your answer.

4. [10 points] **Naive Bayes**

   You have a classification problem with 4 classes. Can you use Naive bayes to solve it? If your answer is yes, explain what changes you would have to make to the representation as well as to the learning algorithm. If you answer is no, explain why not.

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

Let H be a class of hypotheses of VC dimension $VC(H)$, and assume each hypothesis maps to $\{-1, +1\}$. Now consider the class of hypotheses defined as follows:

$$aH = \{sgn(ah) | a \in \Re, h \in H\}$$

Prove that $VC(aH) \geq VC(H)$. Can this inequality be strict? Explain precisely your answer.

6. [10 points] **Kernels**

   Let K be a function defined on strings of letters of length at most 100, such that $K(x, y)$ counts the number of positions in which x and y have the same letter. For example, $K(\text{“grain”}, \text{“crane”}) = 2$, $K(\text{“mom”}, \text{“mechanic”}) = 1$, . Is K a kernel? Prove your answer.

7. [10 points] **Maximum likelihood estimation**

Suppose that we have a data set with one real-valued input an one real-valued output. We want to compute one unknown parameter w given that we assume that y_i is drawn from a normal distribution with $\mu = e^{wx_i}$ and $\sigma = 1$. Write down the log-likelihood of the data (assuming iid instances as usual), and write an algorithm which maximizes the log-likelihood.