Full Name:____

CS 15-681, 15-781 Fall 1998

Midterm

October 27, 1998

Instructions:

- Make sure that your exam is not missing any sheets, then write your name *on every page indicated*.
- This exam is open book, open notes.
- The exam has a maximum score of 85 points. You have 80 minutes.
- Write your answers in the space provided. If you need extra space, use the back of the preceding sheet.
- Write clearly and be concise.

Full Name:______________________________________

Problem 1. (20 points): Concept Learning

Consider the EnjoySport concept learning task defined in Table 2.2 of the textbook.

A. Give a minimum length sequence of training examples that produces the following version space.

```
S:{< ?  Warm   Normal   Strong   Cool     ? >}
G:{< ?    ?    Normal   Strong    ?       ? >}
```

B. How many additional examples are required to make the version space converge to the following target concept?

```
< ?    ?    Normal   Strong    ?       ? >
```

C. Give a bound on the number of examples required to learn the target function if the learner were not trained by a teacher, but instead used an optimal query strategy. More specifically, assume each training example in the sequence is generated by allowing the learner for suggest an instance, and the trainer to provide its classification. (Note the hypothesis space in this problem contains 973 distinct hypotheses.) Assume that the entire sequence of instances must be generated by the learner, and that the target concept might be any hypothesis in H.

D. Suppose that instead of the student generating instance queries to be labeled, the instances were generated at random according to an unknown probability distribution (e.g. nature). Again, a trainer provides the classification for each new instance. Give a bound on the number of examples required to probably (with probability .8) produce a version space that contains only hypotheses with error of 0.1 or less.

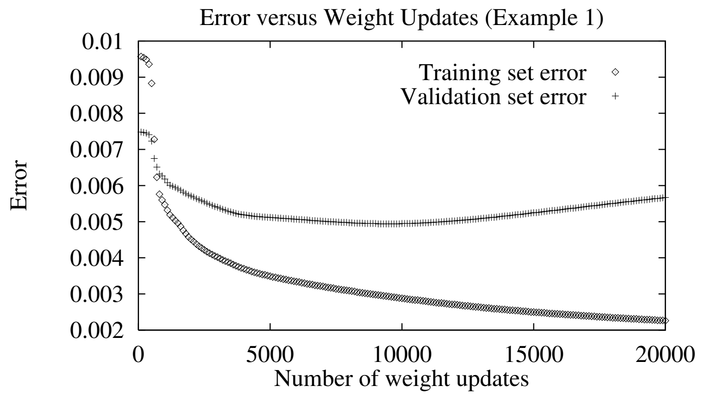
E. Explain in one sentence (with no equations) why you cannot bound the number of examples needed to perfectly learn the target concept (error = 0) under the setting in part D.

Full Name:____________________________________

Problem 2. (30 points): Short Answers

A. We wish to estimate the true error $error_{\mathcal{D}}(h)$ for some decision tree h. We test h on 100 instances drawn at random from the distribution $\mathcal{D}$ and note that it correctly classifies 60 of these instances. Based on this information, give the the most probable value for $error_{\mathcal{D}}(h)$, the standard deviation in this error estimate, and a 95% two-sided confidence interval. You may leave all of these answers as expressions - no need to do the arithmetic.

B. Consider the training error and test error observed as we train a neural network (below) using full (not stochastic) gradient descent. Draw in on this diagram how you would expect both curves to change if we were to rerun this experiment using double the number of training examples.



C. Does Backpropagation descend the error surface that corresponds to the *training set error* or the *true error* of the network? Explain in one sentence.

Full Name:____

D. Name two learning algorithms that search a complete hypothesis space (i.e., all boolean functions over the instances), and that employ an inductive bias favoring simpler hypotheses. Name a third algorithm that can search a complete hypothesis space but uses a different criterion for choosing its hypothesis.

E. Sigmoid threshold units can be used to approximate perceptrons arbitrarily closely. Describe the sigmoid unit weights you would use to very closely approximate the following 3-input perceptron.

$$o(x_1, x_2, x_3) = \begin{cases} 1 & \text{if } 1 + 2x_1 - 4x_2 + 3x_3 > 0 \\ 0 & \text{otherwise.} \end{cases}$$

F. The PAC learning bounds state that "with probability $(1-\delta)$ the learner will succeed in outputting a hypothesis with error at most ϵ." This suggests there is some randomized experiment with the learner that, if repeated n times, would be expected to succeed $n(1-\delta)$ times. What exactly is this experiment?

Full Name:______________________________________

Problem 3. (15 points): VC dimension

Consider the hypothesis space H consisting of arbitrary pairwise conjunctions of perceptrons. In other words, each hypothesis h_i is of the form $p_1 AND p_2$, where p_1 and p_2 are perceptrons. h classifies an instance positive if and only if both p_1 and p_2 classify it as positive.

A. Recall that the decision surface of a single perceptron is a hyperplane. Draw the decision surface corresponding to a pairwise conjunction of 2-input perceptrons.

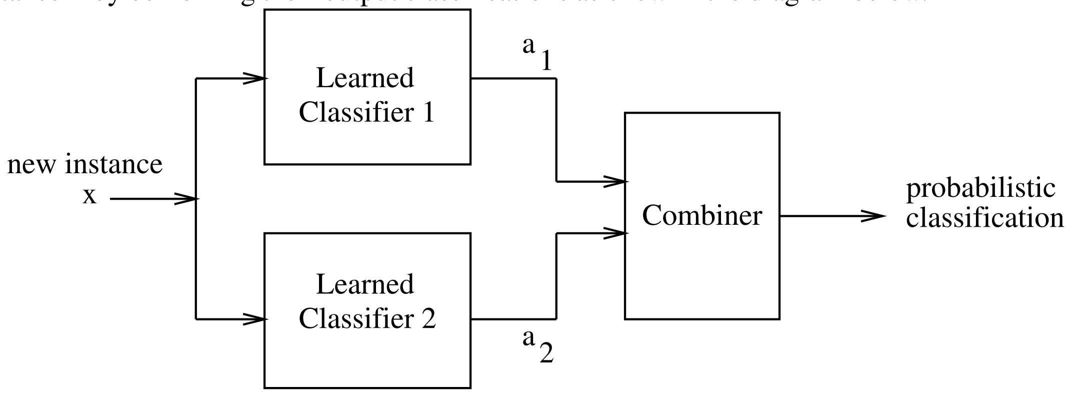
B. What is the VC dimension of this hypothesis space H?

C. How many training examples would be required to assure that with probability at least 0.95 a consistent learner would output a hypothesis whose error is at most 0.1? You may leave your answer as an expression.

Full Name:____

Problem 4. (20 points): Naive Bayes and Combining Classifiers

One common approach to improving classification accuracy is to train multiple classifiers using the same training data, and then combine their outputs in some way. For example, we might use the available training data to learn both a neural network (classifier 1), and a decision tree (classifier 2). We can then classify each new instance x by combining their output classifications as show in the diagram below.



There are numerous ways to implement a Combiner of this type. Here you are asked to implement it using a Naive Bayes classifier.

A. Design a Naive Bayes learner to implement the Combiner. Assume here that the target function has two possible values (v_1 and v_2), and that the output a_i of the ith classifier is simply one of these two values. Therefore, your Combiner may treat a_1 and a_2 as two boolean attributes that describe x (note: your Combiner is not given the description of x itself). Describe the exact formula your Combiner will use to calculate the final classification and its probability:

B. List of all probability terms that your Naive Bayes Combiner will use, and indicate how you would have it learn each term.

C. State any assumptions that your Naive Bayes combiner is making about the relationship between Classifier 1 and Classifier 2.

D. Imagine what would happen if Classifier 1 and Classifier 2 were just two copies of the same neural network. Would your Combiner output a probability that is too low, too high, or just right? Justify your answer in a sentence or two.

E. You notice that one shortcoming of your Naive Bayes Combiner is that it cannot capture the fact that Classifier i is more accurate for some instances x than for other instances. You consider adding x as an additional input to your Combiner, but note that this will certainly violate the Naive Bayes independence assumption. Suggest in a sentence or two a more elaborate probabilistic approach you might use if you want to redesign your Combiner to use x as an input as well.