

**University of Wisconsin-Madison  
Computer Sciences Department**

**CS 760 — Machine Learning**

*Fall 1998*

**Exam**

*(one page of notes and calculators allowed)*

*100 points, 105 minutes*

*December 10, 1998*

Write your answers on these pages and show your work. If you feel that a question is not fully specified, state any assumptions you need to make in order to solve the problem. You may use the backs of these sheets for scratch work. Notice that all questions do not have the same point-value. Divide your time appropriately.

Before starting, write your name on this and all other pages of this exam. Also, make sure your exam contains five (5) problems on eleven (11) pages.

<b>Problem</b>	<b>Score</b>	<b>Max Score</b>
1	_____	36
2	_____	25
3	_____	17
4	_____	12
5	_____	10
Total	_____	100

**1. Learning from Labelled Examples (36 pts)**

Assume you are given three Boolean-valued features -  $F1$ ,  $F2$ , and  $F3$  - and the following training examples. *Show all your work when answering the questions below.*

$F1 = T$	$F2 = T$	$F3 = T$	+
$F1 = T$	$F2 = F$	$F3 = T$	-
$F1 = T$	$F2 = F$	$F3 = F$	+
$F1 = T$	$F2 = T$	$F3 = F$	-

Part A Using *ID3* and its max-gain formula, produce a decision tree that accounts for the training examples. In case of ties, choose the *lowest*-numbered feature.

Part B Discuss how one would apply the *Naive Bayes* algorithm to the above training data. Show how the resulting classifier would categorize the following testset example:

F1 = F      F2 = T      F3 = F      ?

Part C Explain how a *one-nearest neighbor* algorithm would categorize the test example of Part B, given the above training set.

Part D Show how a *perceptron* would be trained using each of the first two (2) examples in the training set once. Initial all weights to 0 and the threshold to 1; set the learning rate to 0.5.

Part E Describe how the *weighted-majority* algorithm could be applied to this data set. Treat each individual feature and its negation as separate prediction algorithms (ie, there are six in total). Assume the examples arrive and are processed in the order of the list above (ie, top-most one first). Break any ties in favor of +.

Part F Assume that you have applied to this task an algorithm that produces a numeric output in the range  $[0, 1]$ . On three "positive" testset examples this algorithm's predictions were:

0.9 0.8 0.4

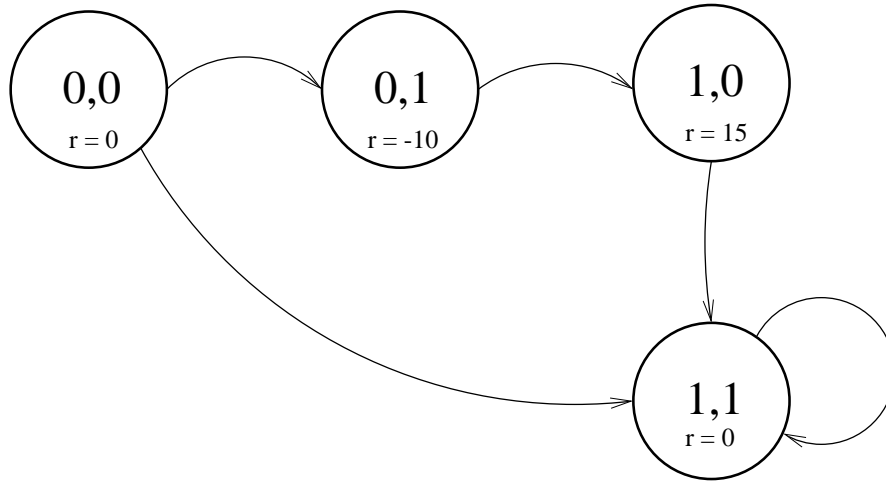
while on three "negative" testset examples the algorithm produced:

0.7 0.3 0.1

Draw an *ROC curve* for these results. Be sure to label your axes and briefly explain your calculations.

**2. Reinforcement Learning (25 pts)**

Imagine an environment that has two Boolean-valued sensors,  $S1$  and  $S2$ , with the legal transitions and immediate rewards shown below. The initial state is always  $\langle S1=0, S2=0 \rangle$  (or  $\langle 0,0 \rangle$  for short).



Part A For what range of values of the *discount rate* (ie,  $\gamma$ ), would a reinforcement learner prefer the *indirect* route from state  $\langle 0,0 \rangle$  to state  $\langle 1,1 \rangle$ ?

Part B Apply the one-step, Q-learning algorithm to this problem, using a *table* to represent your Q-function (assume all entries in the table initially equal 3); let  $\gamma = 0.9$ . Notice that you can "implement" your Q-table by writing the Q values directly on the arcs of the graph above.

Describe the state of the Q-table after each of the first two (2) steps of the learner. For simplicity, always follow the current policy during learning (ie, *no* exploration) and break any ties by moving to the lowest-numbered state (treating  $S1 S2$  as a binary number). Briefly explain below your calculations.

Part C Instead of using a Q-table, imagine that you have chosen to use a nearest-neighbor approach to represent the Q function. Under the assumptions of Part B, what would be the first training example added to your nearest-neighbor memory? Explain your answer.

Part D Imagine now that you have a task that involves  $N$  Boolean-valued features, where all states are reachable from all other states. As a function of  $N$ , what is the ratio of memory needed to (i) store the Q function as a table vs. (ii) the memory needed to represent the Q function as a neural network with one layer of  $0.5xN$  hidden units? Use the standard feedforward neural-network topology that you used in HW3. Compute a numeric value for  $N=50$ .

Part E When incrementally updating estimated Q values, commonly  $\alpha$  times the latest estimate is added to  $(1 - \alpha)$  times the previous estimate. Is such a technique needed in *deterministic* environments when the Q-function is represented in a neural network? Explain your answer.

*Qualitatively* describe any necessary properties of  $\alpha$  (no need to mention that  $\alpha$  is in  $[0,1]$ ).

**3. Computational Learning Theory and Experimental Methodology (17 pts)**

You are assigned the task of learning a "profile" for individual users of a new software package. A profile's input vector contains 32 Boolean-valued measurements, and its output is a single Boolean value that indicates whether or not the user needs help. The specification for this task says that on at least 99% of the users of your software, the profiles that your algorithm learns have to make the correct prediction at least 95% of the time.

Part A     Imagine that you choose to represent profiles as simple conjunctions of a subset of these features (and their negations). Assume that some such conjunction always exists for all users of the software.

Compute how many training examples you need to collect from each user in order to meet this task's specification. You may assume that all data is noise-free.

Part B Next assume that you have created two learning algorithms for this task and have run a 10-fold cross-validation experiment using 1000 labelled examples. The testset accuracies measured on each fold are as follows:

Algorithm 1: 99 95 93 98 97 93 99 98 97 92

Algorithm 2: 98 96 93 97 95 94 99 97 94 94

Can you say, with 95% confidence, that the difference between these two algorithm's accuracies is statistically significant? Show the calculations that justify your answer.

**4. Short Essays (12 pts)**

Briefly explain the importance in machine learning of the following:

---

*boosting*

---

*fitness-proportional reproduction*

---

*MDL hypothesis*

---

*fixed-length feature vectors*

---

**5. Longer Essays (10 pts)**

IF YOU WISH, YOU MAY RIP OFF THIS FINAL QUESTION, TAKE IT HOME WITH YOU, AND RETURN IT BY 5PM NEXT TUESDAY. HOWEVER, DO NOT DISCUSS YOUR ANSWERS WITH ANYONE ELSE UNTIL AFTER THAT DEADLINE (this constraint holds even if you turn in your answer to this question today). You may type your answers on a separate sheet of paper, but do not use more than one side of a normal sheet of paper and use a reasonably large font and wide margins.

Part A Describe what you believe to be the most interesting *algorithm* in machine learning. Briefly justify your answer.

Part B Describe what you believe to be the most important *open issue* in machine learning. Briefly sketch an approach for addressing it.