Machine Learning 15:681

Final Exam, Fall 1996

Carnegie Mellon University

Professors Avrim Blum and Tom Mitchell

Instructions:

1. **WRITE YOUR NAME AT THE TOP OF EACH PAGE.**
2. Answer questions in the space provided. If you need extra space, use the back of the preceding sheet.
3. Short, precise answers will receive the most credit. Long rambling answers are worth fewer points.

Hint: A calculator is *not* needed for this test.

Problem 1: (short answers)

(a) `Backpropagation` and `ID3` are two practically useful machine learning algorithms. For each, describe in one sentence a learning task to which you would *not* apply this algorithm. In a second sentence, describe why the other algorithm would be more appropriate.

(b) Why is explanation-based learning unable to learn to play chess sufficiently well to beat the world champion? Give just one reason, in at most two sentences.

(c) Under what assumptions is the maximum likelihood hypothesis the one that minimizes the sum of squared errors for a dataset?

(d) The `FOIL` algorithm, like many algorithms, can overfit the training data. One approach to avoiding this is to postprune the rules it learns. Describe an alternative method for addressing this problem.

(e) Consider using a naive Bayes classifier to learn the *EnjoySport* target concept from the following data. All the attributes here have exactly two possible values.

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

Exactly how many distinct probabilities must be estimated to train a naive Bayes classifier for this problem? Justify your answer in a sentence or two.

(f) Consider the space of instances consisting of points in the x, y plane, and the set of hypotheses consisting of axis-parallel rectangles in the plane. An axis-parallel rectangle is one whose sides are parallel to the x and y axes. A rectangle hypothesis classifies all interior points as positive examples, and all other points negative. What is the VC dimension of such rectangles in the x, y plane?

Problem 2: (Mistake bound model, conjunctive concepts, neural nets)

(a) Give a mistake-bound learning algorithm for the following situation. Each instance is described by n real-valued features, and the target concept specifies certain $\{\geq, \leq, >, <, =\}$ relations that must hold among pairs of features for an example to be positive. For instance, a typical target concept might be: "positive if $x_1 \geq x_4$ and $x_2 = x_3$ and $x_7 < x_9$".

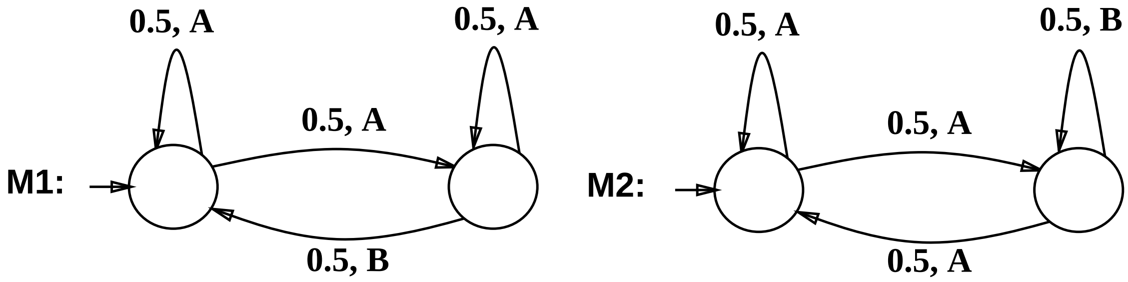
What is your worst-case mistake bound?

(b) Consider a neural network with two inputs, a single output, and no hidden units. That is, the network computes $\sigma(w_0 + w_1 x_1 + w_2 x_2)$. Can this network compute the function "positive if $x_1 = x_2$ and negative otherwise?" (We would threshold the output to get a strict positive/negative prediction.) Explain why or why not in at most two sentences.

(c) Suppose that we want to use a neural network that has no hidden nodes to learn the concepts of part (a). What would be a good input encoding to use for this problem? I.e., what would you give as inputs to this neural network?

Problem 3: (HMMs and Bayesian reasoning)

Consider the following two Hidden Markov Models:



(a) Consider string S = "AAAB". What is $\mathbf{Prob}(S|M1)$? What is $\mathbf{Prob}(S|M2)$?

(b) Suppose our prior probabilities state that apriori, model M2 is 3 times a likely as model M1: i.e., $\mathbf{Prob}(M1) = 1/4$ and $\mathbf{Prob}(M2) = 3/4$. (For instance, these models might represent words, and the string S might represent a sequence of phonemes, and we might have apriori probabilities on which word we think is most likely.)

What are the posterior probabilities of the two models, given string S?

(c) Given the posteriors you calculated in (b), what is the probability that the next letter seen after S will be a "B"?

Problem 4

(a) Give an example of a rule that would have predicted the winner in the last 6 presidential elections, but that you would expect to do poorly on the next presidential election. The last 6 presidential elections were:

Winner	Loser
Carter	Ford
Reagan	Carter
Reagan	Mondale
Bush	Dukakis
Clinton	Bush
Clinton	Dole

(b) Given several hypotheses, one of which has predicted correctly on all past data, and the rest of which have made some mistakes, should one always prefer the rule that predicted correctly? Give a short discussion of this from a *Bayesian* perspective.

(c) Same as (b), but answer from a *PAC-model* perspective.

Problem 5: Combining inductive and analytical learning.

Consider learning the target concept $PlayTennis$, given training examples described by the boolean attributes *Humid, Cool, WeekendDay, FullMoon, SchoolHoliday.*

(a) Draw the neural network created by the `KBANN` algorithm to learn this target concept, given the following domain theory. Include all units and all connections. Do not bother to give the exact weight values, but do indicate which weights are non-negligable.

$$PlayTennis \leftarrow NiceWeather \wedge NotWorking$$

$$NiceWeather \leftarrow Cool \wedge \neg Humid$$
$$NotWorking \leftarrow WeekendDay$$

You are asked to run an experiment comparing the performance of `KBANN` to that of `Backpropagation`. You choose to use a fully connected feedforward network with one hidden layer that contains two hidden units. In your experiment you give both learners the same set of m training examples for the $PlayTennis$ target concept. You then test the accuracy of the two learned networks on a separate set of data.

(b) When m is small, which learned network do you expect to be most accurate? Briefly explain why, including any assumptions you are making.

(c) As $m \to \infty$ do you expect the two learned networks to approach the same level of accuracy? Explain your answer, and state any assumptions you are making.

Problem 6: VS, Bayes Optimal, WeightedMajority

Consider the following algorithms, all of which combine the votes of multiple hypotheses to classify a new instance. Suppose all of these algorithms are applied to the same concept learning task, consider the same hypothesis space H, learn from the same training data D, and are then asked to classify the new instance x.

- The version space `Candidate-Elimination` algorithm, which classifies the new instance x by a simple vote among the members of the version space $VS_{H,D}$.
- A Bayes optimal classifier that considers the set H of hypotheses.
- The `WeightedMajority` algorithm, which treats each hypothesis in H as a separate prediction algorithm.

(a) The `Weighted-Majority` algorithm is defined precisely in Table 7.1 of the textbook. Describe how to instantiate `Weighted-Majority` so that its classification of the new instance x will be identical to that of the `Candidate-Elimination` algorithm for all x.

(b) Describe how to instantiate the Bayes optimal classifier so that its classification of the new instance x will be identical to that of the `Candidate-Elimination` algorithm for all x.

(c) Describe how to instantiate the Bayes optimal classifier so that its classification of the new instance x will be identical to those of `Weighted-Majority` algorithm using $\beta = 0.5$, for all x.