

## Final Exam

*May 4, 2001*

- For any problem that involves any arithmetic you need not give the final answer. Just be sure to put it in a form where all that would be left to do is plug it into a calculator.
  - If you have any questions about what is being asked in a problem, please come ask me to clarify the problem.
1. (10 pts) You are walking in the Central Park and you see a women who looks like the actress Julia Roberts. You decide to take a Bayesian approach to decide if she really is Julia Roberts. Based on what the women looks like you think that there is a 25% chance that she is Julia Roberts, and a 75% chance that she is not.

You approach her and ask her whether she is Julia Roberts and she answers “yes.” You reason as follows. “If she is Julia Roberts, she might not want to admit it. I’d say that if she is Julie Roberts, then there’s a 50% chance she’ll say she is, and a 50% chance she’ll say she’s not. On the other hand, if she’s not Julia Roberts, she might think it’s funny to pretend that she is. I’d say that if she isn’t Julie Roberts, then there’s a 25% chance she’ll say she is, and a 75% chance she’ll say she’s not.”

Based on your estimates of the probabilities, is it more likely that she is Julia Roberts, or that she is not Julia Roberts? Show your work.

2. (25 pts) Consider the following collection of data:

Pizza	Temperature	Crust	Topping	Delicious
1	hot	thick	sausage	+
2	hot	thick	olive	-
3	mild	thick	pineapple	+
4	hot	thin	sausage	-
5	cool	thin	olive	-
6	cool	thin	pineapple	+

USEFUL FACTS:

$$-(1/3) \log_2(1/3) = .52832083$$

$$-(1/2) \log_2(1/2) = .5$$

(a) What is the entropy of this data set? Briefly explain how you got your answer.

(b) Given the decision tree that would be output by ID3. Show your work.

(c) Using your decision tree from (b), how would you classify the example “Temperature=cool, Crust=thin, Topping = sausage”?

(d) Use Naive Bayes to give the likelihood (based on the data given on the last page) that the example “Temperature=cool, Crust=thin, Topping = sausage” is a delicious pizza. Then indicate what prediction Naive Bayes would make. Show your work.

(e) A good measure of distance between two examples is the number of attributes which have different values. Using this measure of distance what classification (+ or -) would be predicted by using 3-NN for the example “Temperature=cool, Crust=thin, Topping = sausage”?  
Be sure to explain your answer.

3. (10 pts) The students in a machine learning course had the following semester averages:

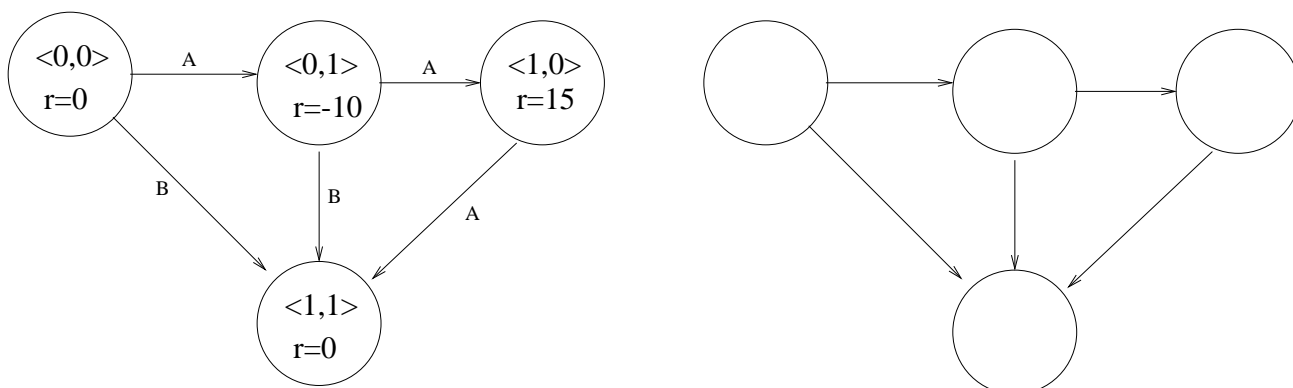
{85, 84, 70, 82, 87, 94, 88, 76, 65, 79, 93, 68, 70, 59, 78, 99, 95, 58, 85, 82, 83, 52}

Only 5 different letter grades,  $\{A, B, C, D, F\}$  can be assigned. Suppose that the instructor begins the semester with the initial expectation that the mean grade for an A to be a 90, for a B to be an 80, for a C to be a 70, for a D to be a 60, and for an F to be a 50. However, the grades assigned are to be determined by finding the best fit with the data under the assumptions that the students receiving each letter grade should have individual grades that are normally distributed (with standard deviation 1).

Describe a method you could use to assign the grades. Give enough detail that I know that you *could* apply it to the above data. You need not actually apply it.

4. (10 pts) Suppose you have 1000 training examples and want to perform 10-fold cross-validation to estimate the generalization error of learning algorithm  $A$ . Describe how you would do this with enough detail that I know you could perform it if I gave you  $A$  and the 1000 training examples.

5. (15 pts) Imagine an environment in which a robot has two Boolean-valued sensors  $S1$  and  $S2$  that define the state of the robot. Consider the MDP shown below where the initial state is always  $\langle S1 = 0, S2 = 0 \rangle$  (or  $\langle 0, 0 \rangle$  for short).



Note that a reward is associated with each state (versus with a reward,action pair). Given a reward and action the immediate reward is that of the state which is reached.

- (a) On the blank graph above, show the  $Q$  and  $V^*$  values for a discount rate of  $\gamma = 0.9$ .  
 (b) What is the optimal policy for  $\gamma = .9$ ?

- (c) For what range of values of the discount rate (i.e.  $\gamma$ ) would the optimal policy for getting from  $\langle 0, 0 \rangle$  to  $\langle 1, 1 \rangle$  be the indirect route defined by always taking action  $A$ ?  
*Hint: Write an expression for the discounted reward for each policy and then you can algebraically determine the requirements on  $\gamma$  for when the indirect route will be optimal.*

6. (10 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 your algorithm learns have to make the correct prediction at least 95% of the time. You are told that some simple conjunction of two of these features (or their or their negations) exists that will always classify the data correctly. You may also assume that the data is noise-free. Describe an efficient learning algorithm for this problem. Be sure to specify exactly (i.e. give a number) how many training examples your algorithm will require and how you will process them.

7. (10 pts) Briefly explain the importance of each of the below items in Machine Learning. You should include how it is applied. No points will be given for just giving a definition or stating what the item is.

(a) VC-dimension

(b) boosting

(c) MDL principle

(d) cross-validation

(e) Bayesian belief network

8. (10 pts) In each of these problems you should describe characteristics of a situation in which one algorithm/approach is better than another. Be sure to not just describe the algorithms or properties about them. For example, saying ID3 is good when you want to pick an attribute that maximizes information gain is not a good answer.

(a) When is a Support Vector Machines (SVM) more appropriate than a Perceptron?

(b) When is Q-learning using a neural network to represent the Q-value estimates more appropriate than using a table to represent the Q-value estimates?

(c) When is the Weighted Majority Algorithm more appropriate than Boosting?

(d) When is a radial-basis network more appropriate than a neural network?

(e) When is Naive-Bayes more appropriate than ID3?

Problem	Points Possible	Points Received
1	10	
2	25	
3	10	
4	10	
5	15	
6	10	
7	10	
8	10	
total	100	