Final Exam Questions.

Computing 882
Simon Fraser University
Spring 2002
Instructor: Oliver Schulte

1. Give decision trees to represent the following Boolean functions:

   (a) $A \wedge \neg B$

   (b) $[A \wedge B] \vee [C \wedge D]$

2. What is overfitting? Explain how noise in the data can lead to overfitting with decision trees.

3. What is overfitting? Explain one general technique that is used to reduce overfitting.

4. Design a two-input perceptron that implements the Boolean function $A \wedge \neg B$. Design a two-layer network of perceptrons that implements $A\ XOR\ B$. Can you design a one-layer perceptron that implements $A\ XOR\ B$?

5. What is an activation function in a neural network? Describe some general properties that an activation function ought to have.

6. What is the relationship between backpropagation and gradient descent?

7. Explain in general terms how gradient descent searches for a minimum of a function, for example a function $f(x)$ of just one variable. Under what assumptions can you apply gradient descent? Is gradient descent guaranteed to find a minimum of $f(x) = x^2$ no matter what the initial guess $\widehat{x}$ is? Is gradient descent guaranteed to find a minimum of $f(x) = \sin x$ no matter what the initial guess $\widehat{x}$ is?

8. Explain in general terms how gradient descent searches for a minimum of a function, for example for a function $f(x)$ of just one variable. Under what assumptions can you apply gradient descent? Why is it important to decrement the step size during the search?

9. State and prove Bayes' Theorem.

10. What is a MAP learner? What is an MLE learner? How are the two different?

11. Recall the dispute between the tobacco industry and the government on the causes of smoking.

    (a) Draw a Bayes net (causal graph) that represents the government's claim that smoking causes lung cancer, including specifying possible conditional probability tables. Write out all the independence relations that hold between the two variables "smoking" and "lung cancer" and various subsets of them.

(b) Draw a Bayes net (causal graph) that represents the industry's claim that some gene is a common cause of smoking and lung cancer. Specify conditional probability tables so that the probability distribution over smoking and lung cancer is the same as in part a). For example, in both graphs the probability of smoking given lung cancer should be the same. Write out all the independence relations that hold between the three variables "smoking", "lung cancer" and various subsets of them.

(c) Let's add a variable "tar build up". The government expands its model so that smoking causes lung cancer indirectly through tar build up. Draw the corresponding Bayes diagram and write out the conditional independence relations that obtain between the variables "smoking", "tar build up" and "lung cancer" (you don't have to write out the probability tables).

(d) Suppose a statistician working for the tobacco industry proposes that the hidden gene is a common cause of tar build up and smoking, whereas smoking does not cause tar build up. The gene does not cause lung cancer, but tar build up does. Draw a causal graph representing these claims. Are the conditional independence relations among the variables "smoking", "tar build up" and "lung cancer" the same as in part c?

12. Consider the rules "if $a_1 = T \wedge a_2 = F \wedge a_3 = T$ then $c = T$" and "if $a_2 = T$ then $c = F$".

(a) Explain how you can represent each of these rules with a bit string.

(b) Show how you could produce a bit string representing a new rule with a crossover operation.

13. What is the role of the crossover operation in genetic algorithms?

14. Explain the difference between supervised and unsupervised learning. How is reinforcement learning an example of unsupervised learning? Why is decision tree induction an example of supervised learning?

15. Consider a grid world with six adjacent squares and one goal state in the bottom middle cell.

2	3	4
1	G	5

. Let the set of states be $\{1, 2, 3, 4, 5, G\}$ corresponding to the squares. Possible actions are moves from any square 1,2,3,4,5 to any adjacent square. The goal square is absorbing. Any transition into the goal state results in a reward of 10; every other kind of transition carries a reward of 0. Give the V* value for every state in this grid world. Give the $Q(s, a)$ value for every transition. Finally, show an optimal policy. Use $\gamma = 0.8$.

16. Consider the problem of the pigeon pecking to get food. We train the pigeon in the following "Skinner box". The pigeon sees a light which changes colours; the possible colours are green, yellow, red. The apparatus works as follows.

(a) The pigeon gets a corn every time it pecks when the light is green.

(b) The pigeon gets nothing if it pecks when the light is red.

(c) The pigeon gets a corn if the light is yellow, the pigeon pecks and the light turns green..

(d) The pigeon pigeon gets nothing if the light is yellow, the pigeon pecks and the light turns red.

(e) From green or red, the apparatus always turns yellow. From yellow, it turns green 70% of the time and red 30% of the time. Thus the colour of the apparatus is independent of the pigeon's actions.

If the pigeon pecks and gets a corn, it amounts to 1 unit of reward. If the pigeon does nothing, its reward is 0. If the pigeon pecks and doesn't get a corn, it amounts to a reward of -1 (for the effort of pecking). Assume that the pigeon's discount factor is 0.8. Model this problem as a Markov decision process. Describe an optimal policy for the pigeon, and find the value function V^* associated with an optimal policy.

17. Consider the natural learning rule in the grue problem (guess that all emeralds are green until you see a blue one).

(a) How many mistakes does this rule make in the worst case?

(b) Give an argument for why every other learning rule makes more than one mistake in the worst case.

(c) Consider the grue problem as a concept learning problem, as follows: The instances are natural numbers 1, ..., n, ... The sample $\langle 1, + \rangle, \langle 3, - \rangle$, for example, represents the observation that the first emerald is green (+) and the third blue (-). The concept space C is the set of concepts corresponding to "all emeralds are green", "all emeralds are grue(1)", ..., "all emeralds are grue(n), ...". What is the VC dimension of this concept space C? (Explain your answer in terms of shattered subsets). How does this fact illustrate the general relationship between VC dimension and mistake bound?

(d) Based on the VC dimension from part (c), for given ε, δ, how many green emeralds would a learner have to observe until it would have confidence δ that "all emeralds are green" is correct up to a probability of ε? After giving the general formula, find the specific required sample size for $\varepsilon = 10\%$, $\delta = 80\%$.

18. Let X be an instance space containing all points in the x, y plane. Give the VC dimension of the following hypothesis spaces.

(a) H = the set of all rectangles in the plane. How many samples does a learner require until it would have confidence 80% that the "most specific" rectangle fitting the data is correct up to a probability of 10%? (Explain your answer.)

(b) H_c = the set of all cirecles in the plane. How many samples does a learner require until it would have confidence 80% that the a circle fitting the data is correct up to a probability of 10%? (Explain your answer.)