C540-1 ANSWER SHEET

First Name _____ Last Name _____ Email _____

1	2
3	4
5	6
7	8
9	10
11	12
13	14
15	16
17	18
19	20

Final Examination CS540-1: Introduction to Artificial Intelligence Fall 2012

20 questions, 5 points each

INSTRUCTIONS: WRITE YOUR ANSWERS ON THE ANSWER SHEET. WE WILL NOT GRADE ANSWERS ON OTHER PAGES. WRITE DOWN THE ANSWERS ONLY -- DO NOT INCLUDE INTERMEDIATE STEPS OR DERIVATIONS. BE SURE TO INCLUDE YOUR NAME AND EMAIL ON THE ANSWER SHEET, TOO.

1. (Machine learning basics) True or False:

   a. A binary classifier with accuracy 0.7 is more useful than a binary classifier with accuracy 0.1.

   b. In a feature vector, all feature values must be non-negative.

   c. In a training set, no two items can have the same feature vector.

   d. After a classifier has been learned on a training set, one can randomly sample 20% of the training set to form a test set, and use the test set to evaluate the accuracy of the classifier.

   e. One can copy a training set 10 times to form a larger training set in order to learn a better classifier.

   All false.

   - One can negate the prediction of the 0.1 classifier, which will have accuracy 0.9
   - Feature values can be any number; in fact they don't have to be numerical
   - Multiple items can have the same feature vector. Recall height and weight of a person.
   - One cannot re-use even part of the training set for evaluation.
   - There is no new information.

2. (Hierarchical Clustering) Consider six points in 2D, some of them overlap. We start with three clusters (rather than from scratch):

   Cluster 1 = (0,0), (1,1), (2,2)

   Cluster 2 = (0,1), (1,0)

   Cluster 3= (0,0)

   With Euclidean distance and complete linkage, <u>which clusters</u> are merged in the next step of hierarchical clustering?

Complete linkage measures the longest distance between points in two clusters.

D(cluster1, cluster2)=Euclidean((2,2),(0,1)) = sqrt(5)

D(cluster1, cluster3)=Euclidean((2,2),(0,0)) = sqrt(8)

D(cluster2, cluster3)=Euclidean((0,1),(0,0))=1

Clusters 2 and 3 will be merged.

3. (K-means clustering) Consider 30 evenly spaced points (with integer features 1 to 30) in 1D and 10 clusters. The current cluster assignments are: points 1,2,3 belong to cluster 1, points 4,5,6 belong to cluster 2, and so on until points 28,29,30 belong to cluster 10. The current cluster centers are c1=1, c2=4, c3=7,…, c10=28. Note these centers are not optimal. Run k-means until convergence. What is the reduction of distortion?

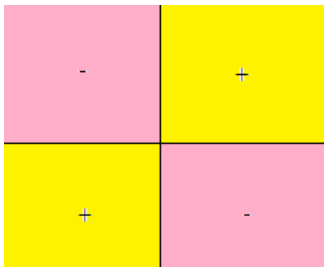
The old distortion: 10*((1-1)^2 + (2-1)^2 +(3-1)^2) = 10*5=50

The new cluster centers will be 2, 5, 8, …, 29. The new distortion is 10*((1-2)^2 + (2-2)^2 + (3-2)^2)=20

The reduction of distortion is 50-20=30.

4. (kNN) Consider points in 2D and binary labels. Given the following training data: x1=(1,1), y1=1, x2=(-1,1), y2=0, x3=(-1,-1), y3=1, x4=(1,-1), y4=0, and use Manhattan distance, how will 1NN classify any points in 2D? Answer this question by drawing the decision boundaries and clearly mark the predicted label of each region.

See figure:



5. (Mutual information) The weather outside can be sunny, raining, snowing, or hailing with equal probability but you are in a windowless classroom and don't know which one. Your professor comes in complaining about car windows being smashed by hails. How much information do you gain on the weather?

H(weather)=H(1/4,1/4,1/4,1/4)=2 bits. H(weather | prof's car smashed by hails)=0 bits. The mutual information or information gain is 2-0=2 bits.

6. (Decision trees) Consider a decision tree with d levels (a tree with a single leaf node and no children has 1 level), and each internal node has b children. The tree is complete, meaning that all leaves are at level d. What is the minimum number of total training examples so that each leaf can contain at least 10 training examples?

We first compute the number of leaves: there are b^(d-1). Thus we need at least 10*b^(d-1) training examples.

7. (SVM) Recall a linear SVM with slack variables has the objective function $\frac{1}{2}W'W + C\sum_i \varepsilon_i$. What happens to W when the trade-off parameter C=0?

When C=0, the slack variables epsilon's can be anything. This makes all the margin constraints yi (W'Xi+b) >= 1-epsilon_i effectively disappear: No matter what W and b is, one can always find some epsilon to make the inequality hold. Then, the objective can be trivially minimized by W=0. This is what W will be.

8. (Perceptron) Consider a single sigmoid perceptron with bias weight w0=1, a single input x1 with weight w1= -1 (MINUS ONE), and the sigmoid activation function g(h)=1/(1+exp(-h)). For what input x1 does the perceptron output value y (assume 0<y<1)?

g(h)=y

1+exp(-h)=1/y

exp(-h)=1/y-1

h= -log(1/y – 1)

h=1-x1

x1-1=log(1/y-1) ➔ x1=log(1/y -1)+1

9. (Probability) If P(A|B) is three times the value of P(B|A), and P(A)=0.123, what is P(B)?

P(B)P(A|B)=P(A)P(B|A)

P(B)=P(A) P(B|A)/P(A|B)

Since P(A|B)=3 P(B|A), we have P(B)=P(A)/3=0.041.

10. (Independence) Tom flipped a coin of unknown bias 100 time under the same conditions. He told you that trials 2 to 90 are all heads, and trials 91 to 99 are all tails. But he didn't tell you the outcome of trail 1 or trial 100. Given what he told you, is trial 1 or trial 100 more likely to be heads?

Neither. Both trials are equally likely to be heads.

11. (Bayesian network) Given the following Bayesian network A➔B➔C with binary random variables, and the CPTs
    P(A)=0.3
    P(B|A)=0.4, P(B|~A)=0.5
    P(C|B)=0.2, P(C|~B)=0.1
    Compute the joint probability P(~A, ~B, ~C). "~" stands for negation.

    P(~A,~B,~C)=P(~A)P(~B|~A)P(~C|~B)=(1-.3) (1-.5) (1-.1)=.7*.5*.9=0.315

12. (D-separation) Consider a Bayesian network with four nodes A,B,C,D and the following edges: A➔B, B➔C, B➔D. Is A and C conditionally independent given D?
    The only path is A➔B➔C head-to-tail. B is not observed, so it does not block the path (the fact that we observe a descendent of B doesn't matter for head-to-tail). Therefore, A and C are NOT conditionally independent given D. Consider A=(x,y), B=A, C=x, D=y.

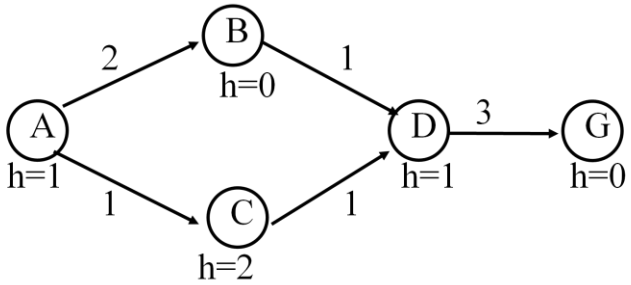
13. (Uninformed search) Say your search problem has a special structure: the state space form a chain s1→s2→s3→…→sn, where s1 is the initial state and sn is the only goal. All states are distinct, all edges have the same cost, and it is possible to compute the predecessor function. Which search algorithm should you avoid: BFS, DFS, uniform cost search, iterative deepening, or bidirectional search?

    all these algorithms can find the goal with similar space requirement (1 or 2). The difference is in time complexity: all are O(n) except for iterative deepening, which is O(n^2). We should avoid iterative deepening. This seems counterintuitive, but is caused by the very small branching factor.

14. (Admissible heuristic function for A*) If h(s) and i(s) are both admissible heuristic functions, what values of p will guarantee that $p * h(s) + (1 - p) * i(s)$ is also an admissible heuristic function?

    p has to be no bigger than 1. Otherwise, p*h(s) when h(s)=H is the true cost will not be admissible. The same argument on i(s) leads to p no smaller than zero. Then, ph+(1-p)i <= pH+(1-p)H=H. So, p in [0,1]. One can also use a convex combination argument.

15. (A* search) List the states in the order of being examined by A* search. Show repeats if any.

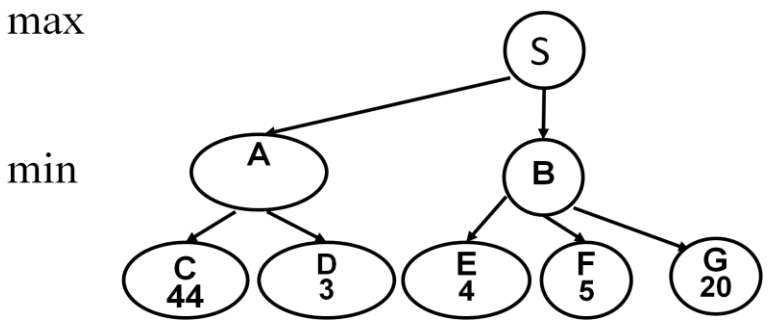


    A(0+1), {B(2+0,A), C(1+2,A)}
    B(2+0,A) {C(1+2,A), D(3+1,B)}
    C(1+2,A) {D(3+1,B), D'(2+1,C)}
    D'(2+1,C) {D(3+1,B), G(5+0,D')}
    D(3+1,B) {G(5+0,D'), G(6+0,D)}
    G(5+0,D')
    The states are A,B,C,D,D,G.

16. (Game tree) Consider a variant of the II-nim game with one stick in one pile and two sticks in the other pile. Max plays first. Recall the rule says each player can take any positive number of sticks from a single pile. Whoever takes the last stick loses. The value for Max is 1 if he wins, and -1 if he loses. What is the game theoretical value of this game?

    The game tree is the subtree of the II-nim game in lecture slides, except that min and max switch order. The game theoretical value is 1 because Max can win by first taking all two sticks from the 2nd pile.

17. (Alpha-beta pruning) Which branch will be alpha-beta pruned in the following game tree?



No alpha-beta pruning happens.

18. (Matrix normal form) What is the game theoretical value of the following matrix normal form game?

	Min-I	Min-II	Min-III	Min-IV
Max-I	3	3	5	5
Max-II	2	2	2	2
Max-III	4	3	4	3

we can take the min of each row, then the max of the results max(3,2,3)=3

or the max of each column, then min(4,3,5,5)=3

19. (Logic) Given the following knowledge base KB:
P
$P \vee Q$
is the query Q entailed by KB?

No. When P=true and Q=false, KB is true but the query is false.

20. (Markov Model) Let there be three states A, B, C (i.e., recall the envelopes in class). We start from state A. Let the transition probabilities be
P(A|A)=0, P(B|A)=P(C|A)=1/2
P(A|B)=P(B|B)=0, P(C|B)=1
P(A|C)=P(B|C)=0, P(C|C)=1
Recall P(x | y) means the probability of going from y to x. After 100 steps, what is the probability that we will be at state A, B, C, respectively?

This Markov process starts at A. It will go to B or C with equal probability after 1 step. However, after 2 steps it will definitely be at C. It will stay at C thereafter. So the answer is 0, 0, 1.