Final Examination CS540-1: Introduction to Artificial Intelligence

December 20, 2010. (20 questions, 5 points each)

LAST NAME: __SOLUTION__________________ **FIRST NAME:**______________________________

EMAIL : ____________________________________

1. (CAPTCHA and Probability) A dog-vs-cat CAPTCHA displays *N* images on the screen. Each image has either a dog or a cat in it. A valid user needs to click on all the dog images but none of the cat images in order to pass the CAPTCHA. Now imagine a robot who cannot distinguish dog vs. cat images. For each image, it has a 50% chance of clicking on that image. What is the probability that the robot will pass the CAPTCHA?

   The robot has a ½ chance of succeeding at each image regardless of its content. The trials are independent. The overall probability of success is therefore 2^{-N}.

2. (Bidirectional Search) Consider a state space where the initial state is the number 1 and each state *k* has two successors: numbers 2*k* and 2*k*+1. Suppose the goal state is 2010.
   a. What is the branching factor in each direction of the bidirectional search?
   forward: 2 , backward: 1
   b. Write down a solution to this search problem.
   1, 3, 7, 15, 31, 62, 125, 251, 502, 1005, 2010

3. (Iterative Deepening) Imagine a world where each person has *b* friends. Alice and Bob are *d* "friendship links" away (i.e., if *d*=1, Alice and Bob are friends; if *d*=2, there is a third person X such that Alice and X are friends, and Bob and X are friends; and so on). Imagine an iterative deepening algorithm that has access to the friendship links. The algorithm starts at Alice and the goal is to find Bob. <u>Do not use big-O notation</u>, instead give precise expressions below:
   a. In the worst case, how much stack memory is required (one unit of memory stores one person)?
   (b-1)d
   b. In the worst case, how many people the algorithm needs to visit (including Alice and Bob)?
   $1+b+b^2+\ldots+b^d=(b^{d+1}-1)/(b-1)$

4. (A* search) Two friends live in different cities on a map. On each turn, we can simultaneously move each friend to a neighboring city on the map or stay in the same city. The amount of time needed to move from city *i* to neighbor *j* is equal to the road distance *D(i,j)* between the cities. But on each turn the friend that arrives first must wait until the other one arrives (and calls the first on her cellphone) before the next turn can begin. We want the two friends to meet as quickly as possible.
   a. Let *S(i,j)* be the straight-line distance between cities *i* and *j*. Let one friend be at city *i* and the other friend be at city *k*. Circle the heuristic functions below that are admissible.

i. $\min_{\{j \text{ in all cities}\}}$ (max($D(i,j)$, $D(k,j)$)) no, it might be faster to go through an intermediate city.

ii. $\min_{\{j \text{ in all cities}\}}$ (max($S(i,j)$, $S(k,j)$)) yes, this is physics

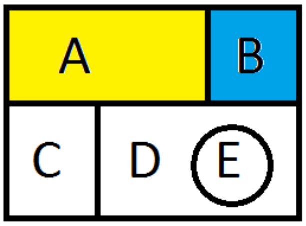
b. Are there completely connected maps for which no solution exists?
no, they can always meet

c. Are there maps in which all solutions require one friend to visit the same city twice?
no, he might as well stay in that city.

5. (Gradient Descent) Let $X=(x_1, \ldots, x_d)$ be a d-dimensional real-valued feature vector. We want to minimize $f(X)=x_1+x_2^2 + x_3^3 + \ldots + x_d^d$ using gradient descent. Let the step size be $\alpha=0.1$. If we start at X=(1,1,…,1) the all-one vector, which X will we be at after <u>one</u> iteration of gradient descent?
$X=X-\alpha\nabla$. The i-th dimension of the gradient is $i(x_i)^{i-1}$. So the new X=(1-0.1, 1-0.2, …, 1-0.1*d)

6. (Game Playing) Two players take turns to work on an 8-puzzle problem (A makes the first move, B makes the second move, A makes the third move, and so on). Whoever solves it first wins. <u>In one sentence</u>, describe what you think will happen in this game.
A wins if in one move he can solve the 8-puzzle. Otherwise, it is an infinite loop and nobody wins.

7. (Arc Consistency) Consider the following map with five areas A (colored yellow), B (colored blue), C, D, E. No two areas that share part of an edge can have the same color.



a. Suppose the candidate colors are yellow, blue, and red. What is the output of the AC-3 algorithm?
A={Y}, B={B}, C={B}, D={R}, E={BY}

b. Repeat the question but suppose the candidate colors are yellow, blue, red, and black.
A={Y}, B={B}, C={BRK}, D={RK}, E={BRKY}

8. (Resolution) Given the following propositional knowledge base:
$A \Leftrightarrow B$
use resolution to prove the query:
$\neg A \lor B$
CNF: KB: $\neg A \lor B$, $\neg B \lor A$; negative query: $\neg B$, A
$\neg A \lor B, \neg B \rightarrow \neg A, A \rightarrow$ empty

9. (Unification) For each pair of atomic sentences, give the most general unifier if it exists, otherwise say "fail":

a. R(A, x), R(y, z) y/A, x/z

b. P(A, B, B), P(x, y, z) x/A, y/B, z/B

c. Q(*y*, G(A, B)), Q(G(*x*,*x*), *y*) fail

d. Older(Father(*y*), *y*), Older(Father(*x*), John) x/y, y/John

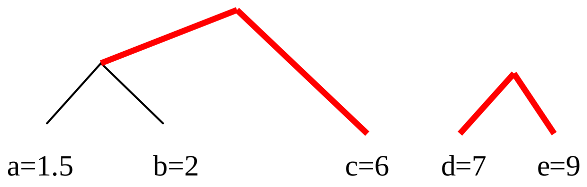
e. Knows(Father(y),y), Knows(x,x) fail

10. (Entropy) Running from You-Know-Who, Harry enters the CS building on the 1st floor. He flips a fair coin: if it is heads he hides in room 1325, otherwise he climbs to the 2nd floor. In that case he flips the coin again, if it is heads he hides in CSL, otherwise he climbs to the 3rd floor. In that case he flips the coin yet again, if it is heads he hides in 3331, otherwise he hides in the Men's room. What is the entropy of Harry's location?
    p=(1/2, ¼, 1/8, 1/8), H(p)=1/2*1+1/4*2+1/8*3+1/8*3=1.75 bits

11. (Decision Trees) There are 100 parrots. They have either a red beak or a black beak. They can either talk or not. Complete the two cells in the following table so that the mutual information (i.e., information gain) between "Beak" and "Talk" is zero:

Number of parrots	Beak	Talk
10	Red	Yes
30 or 15	Red	No
15 or 30	Black	Yes
45	Black	No

12. (Clustering) There are five points in one-dimensional space: a=1.5, b=2, c=6, d=7, e=9. Perform Hierarchical Agglomerative Clustering with Single Linkage, with the following extra constraint: at no time should *c* and *d* be in the same cluster. Complete the resulting clustering tree diagram (i.e., the dendrogram).



13. (kNN) True or False:

    a. If A is among B's k-nearest-neighbors, then B is among A's k-nearest-neighbors.
    False

    b. On a training set consisting of 1000 positive items and 1000 negative items, it is possible for 1NN to have training set accuracy 0 (i.e., it can completely fail on the training set).
    True

    c. kNN results may change if we multiple 0.1 to each dimension in every item's feature vector (assuming we use Euclidean distance).
    False

14. (Support Vector Machines) Consider a small dataset with four points, where each point is in 2D:
    (x11=0, x12=0), y1=0
    (x21=0, x22=1), y2=1

$(x31=1, x32=0), y3=1$

$(x41=1, x42=1), y4=0$

a. Can a linear SVM perfectly classify this dataset? No. XOR not linearly separable

b. What if we map each feature vector $X=(x1, x2)$ into $\phi(X)=(2*x1, 2*x2, -x1-x2)$?
still no, this is a linear mapping.

c. In general, is there a set of three 2-dimensional points such that, no matter what binary label we give to each point, a linear SVM can perfectly classify the resulting dataset?
yes. Any three points that are not co-linear will do.

d. Same as above, but with four instead of three points.
no. this is easy to see as in the XOR example.

e. Same as above, but with five points.
no. this is strictly harder than four points.

15. (Bayes Network) Consider a Bayes network $A \rightarrow B \leftarrow C$ with binary variables and CPTs:

$P(A=T)=1/4$

$P(B=T \mid A=T, C=T)=1/3$

$P(B=T \mid A=T, C=F)=1/\pi$

$P(B=T \mid A=F, C=T)=1/42$

$P(B=T \mid A=F, C=F)=2/(1+\sqrt{5})$

$P(C=T)=0.540$

Compute $P(A=F \mid C=F)$.

A and C are independent. $P(A=F|C=F)=P(A=F)=1-1/4=3/4$

16. (Speech Recognition) Traditional speech recognition can be posed as a probabilistic inference problem: given acoustic signal A, the task is to find a sentence W such that

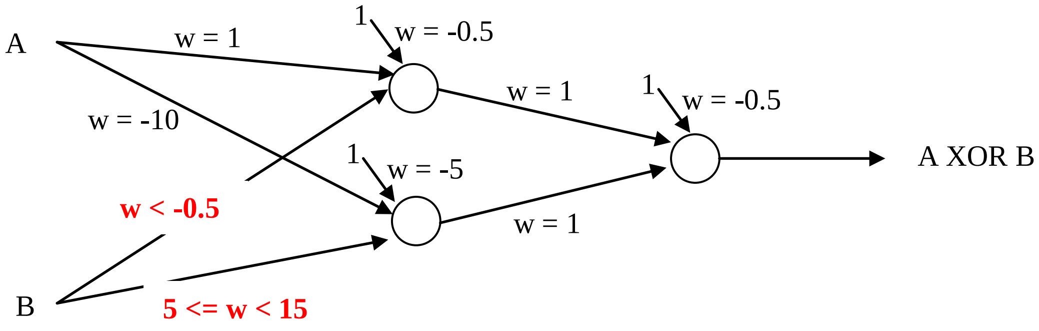
$W^* = \text{argmax}_W P(W \mid A) = \text{argmax}_W P(A \mid W) P(W)$ (1)

where $P(A|W)$ is the acoustic model and $P(W)$ is the language model. In light of the McGurk effect, video signal V of the speaker's face is also helpful in speech recognition. In one line, write down how you would modify equation (1) to incorporate both acoustic and video signals for speech recognition. In another line, briefly explain the components in English.

$W^* = \text{argmax}_W P(W \mid A,V) = \text{argmax}_W P(A,V \mid W) P(W) = \text{argmax}_W P(A|W) P(V|W) P(W)$

Same AM and LM, but we have $P(V|W)$ as a "video model" now.

17. (Neural Networks) Fill in the missing weights below so that it computes A XOR B. Both A and B takes value 0 or 1, and the perceptrons are Linear Threshold Units (LTUs).



18. (Probabilistic Inference) There are three biased coins in a bag. The probability of generating Heads by the coins is: coin A 0.2, coin B 0.6, coin C 0.8, respectively. I drew one coin from the bag with equal probability. I flipped it ten times and the outcome was 3 Tails and 7 Heads. Which coin did I most likely draw?
argmax_c P(c|7H, 3T)=argmax P(7H,3T|c)P(c) = argmax P(7H,3T|c)
P(7H,3T|c=A)=0.2^7 * 0.8^3 = 6.55e-6
P(7H,3T|c=B)=0.6^7 * 0.4^3 = 0.00179
P(7H,3T|c=C)=0.8^7 * 0.2^3 = 0.00168 ➔ I most likely picked coin B.

19. (Machine Learning) We have a biased coin with probability 0.8 of producing Heads.
    a. We create a predictor as follows: generate a random number uniformly distributed in (0,1). If the random number is less than 0.8 we predict Heads, otherwise we predict Tails. What is this predictor's accuracy in predicting the coin's outcomes?
    0.8*0.8+0.2*0.2=0.68
    b. We create another predictor which always predicts Heads. What is this second predictor's accuracy in predicting the coin's outcomes?
    0.8

20. (kNN) You have a data set with 50 positive items and 50 negative items. You perform a so-called "leave-one-out" procedure: for each i=1,2,…,100, learn a separate kNN classifier on all items except the i-th one, and compute that kNN's accuracy in predicting the i-th item. Call this accuracy a_i. The leave-one-out accuracy is defined to be the average of a_1, a_2, …, a_{100}. What is the leave-one-out accuracy when k=99?
Because each 99-NN has a majority vote that's the opposite of the left out item, all the ai's will be zero. So the leave-one-out accuracy with 99NN is zero.