CS540 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 2013
20 questions, 5 points each

INSTRUCTIONS: WRITE YOUR ANSWERS ON THE ANSWER SHEET. WE WILL NOT GRADE ANSWERS ON OTHER PAGES. ON THE ANSWER SHEET 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. (Smoothing) You have a dictionary with 1,000,000 entries (these are called *word types*). You want to estimate the probability $p(w)$ for each word type w in your dictionary. NSA sends you the count c_w, i.e. the number of times w appearing in all human online communications in 2012, for all the w's in your dictionary. The total count $\sum_w c_w$ is 10^{12}, and $c_{\text{“olinguito”}} = 3$. Using add-one smoothing, write down the formula for $p(\text{“olinguito”})$ (your answer should not contain variables, but do not numerically compute the decimal number either).

2. (HAC) In 1D there are three clusters: (1, 3, 7), (4, 10), (12). Will single linkage HAC and complete linkage HAC merge the same two clusters next?

3. (Search) Suppose the state space has 10^{10} states that form a tree. What tree shape makes iterative deepening realize that a goal state does not exist as quickly as possible? Describe the tree in one sentence.

4. (SVM) Recall the margin of a point $\mathbf{x}$ is defined as $\mathbf{w}^\top\mathbf{x}+b$. Suppose $\mathbf{w} = (1, 2, 3)^\top, b = 0.99$, what binary label will SVM assign to the point $\mathbf{x} = (1, 1, -1)^\top$?

5. (kNN) Consider binary classification in 2D where the intended label of a point $\mathbf{x} = (x_1, x_2)$ is positive if $x_1 \geq x_2$ and negative otherwise. Let the training set be all points of the form $\mathbf{x} = (a, 2b)$ where a, b are integers. Each training item has the correct label that follows the rule above. With a 1NN classifier, what is the label of test point $\mathbf{x} = (1.4, 1.1)$?

6. (A^* search) True or false: If on every state the heuristic function $h()$ over-estimates (i.e., is strictly larger than) the true cost from that state to the optimal goal by 100, then A^* search with this $h()$ is still guaranteed to find the optimal goal.

7. (Kernel) If $K(\mathbf{x}, \mathbf{x}')$ is a kernel whose induced feature representation is $\phi()$, and $G(\mathbf{x}, \mathbf{x}')$ is a kernel whose induced feature representation is $\psi()$, then $H(\mathbf{x}, \mathbf{x}') = aK(\mathbf{x}, \mathbf{x}') + bG(\mathbf{x}, \mathbf{x}')$ is a valid kernel if $a > 0, b > 0$. Write down the feature representation induced by H.

8. (Neural Nets) In a three-layer neural network, the first layer contains 1,000 sigmoid units, the second layer contains 500 units, and the output layer contains 1,000 units. The input $\mathbf{x} \in \mathbb{R}^{1000}$. How many weights does this neural network have?

9. (Machine learning) True or false: if you have a training set and a separate test set, it is better if you remove any test item (x, y) from the test set if its feature vector x has appeared in the training set. That is, you should remove test item (x, y) if there is at

least one training item (x, y') with the same x regardless of its label y'. If you don't, your computed test accuracy will not be accurate.

10. (Gradient descent) Let $f(x) = 5x^2$. Gradient descent iteratively performs $x \leftarrow x - \eta f'(x)$ where $f'(x)$ is the derivative of f at x. What range of η guarantees that $x \to 0$?

11. (Probability) It is known that $P(A) + P(B) = 0.5$, and $\frac{P(A|B)}{P(B|A)} = \frac{1}{4}$. What is $P(A)$?

12. (Entropy) Statistically, December 18 is the cloudiest day of the year in Madison, Wisconsin. Your professor is not making this up. On that day, the sky is overcast, mostly cloudy, or partly cloudy $\frac{2}{3}$ of the time ($C = 0$), and clear or mostly clear $\frac{1}{3}$ of the time ($C = 1$). What is the entropy of the binary random variable C? Keep the *log*.

13. (Game tree) Consider a variant of the II-nim game. There are two piles, each pile has two sticks. A player can take one or two sticks from a single pile, *or she may take two sticks, one from each pile (when available)*. The player who takes the last stick loses. Let the game value be 1 if the first player wins. What is the game theoretical value of this game?

14. (K-means) Consider K-means on a data set in 1D where, as initialization, 1, 3, 5, 7, 9 are assigned to cluster I, and 2, 4, 6, 8, 10 are assigned to cluster II. What will the cluster assignment be in the next iteration of K-means?

15. (Independence) Alice rolls a six-sided die. If Alice's outcome is 1 or 2, Bob flips a fair coin; If Alice's outcome is 3, 4, 5, or 6, Bob flips a biased coin with head probability 0.8. If Bob's outcome is heads, Chuck draws a card randomly from a Poker deck where the four kings have been removed beforehand; If Bob's outcome is tails, Chuck draws a card randomly from another Poker deck where the four aces have been removed beforehand. Finally, if Chuck gets a queen David raises a red flag; otherwise David raises a green flag. Are Alice and David's outcomes independent?

16. ($\alpha - \beta$ pruning) Standard $\alpha - \beta$ pruning starts by calling Max-Value(root, $-\infty$, ∞). Now consider a zero-sum game where all leaves take values in $\{-1, 0, 1\}$. True or false: in this game one can start by calling Max-Value(root, -1, 1) and still perform the correct $\alpha - \beta$ pruning.

17. (Decision tree) A hospital trains a decision tree to predict if any given patient has technophobia or not. The training set consists of 1,000,000 patients. The features include the gender, age, social security number, blood type, and 12 medical tests. All features are discrete. The labels are binary. The decision tree is not pruned. What is the training set accuracy of the decision tree?

18. (Directed graphical model) True or false: for any directed graphical model with at least one directed edge $v_i \rightarrow v_j$, one can always create another directed graphical model which encodes exactly the same joint probability distribution but with that edge flipped: $v_i \leftarrow v_j$ (there is no requirement to keep the rest of the graphical model the same).

19. (Dominating strategies) Given the following matrix normal form where the first number is the payoff to A and the second to B, what is the pure strategy that the two rational players will adopt?

	B-I	B-II	B-III	B-IV
A-I	9, 7	3, 8	3, 8	1, 1
A-II	2, 9	4, 9	5, 5	2, 8
A-III	4, 0	6, 1	9, 1	0, 0
A-IV	2, 5	3, 6	5, 7	1, 8

20. (Search state space) There are 11 lights in a row. The initial state is (□ = on, ■ = off):

□■□■□■□■□■□

A valid move finds two adjacent lights where one is on and the other is off, and switches them while keeping all other lights the same. That is, locally you may do □■ → ■□, or ■□ → □■. What is the smallest number of moves to reach the goal state:

□□□□□■■■■■■