Name ____ Email ______________________

CS540-1 F17 final exam. If it is a multiple choice question, circle ONE answer; otherwise write the answer (no intermediate steps please).

1. (Zipf's law) According to Zipf's law, if a word w_1 ranks 3rd and w_2 ranks 7th, what is the ratio f_1/f_2 between the frequency of the two words?
   A: Since $fr = c$ for some constant c, we have $f_1 = c/3$, $f_2 = c/7$, thus $f_1/f_2 = 7/3$.

2. (Optimization) Alice, Bob and Cindy go to the same school and live on a straight street lined with evenly spaced telephone poles. Alice's house is at the 2nd pole, Bob's at the 4th pole, Cindy's at the 9th pole. Where should the school set up a school bus stop so that the sum of distances (from house to bus stop) walked by the three students is minimized?
   A: The problem is $\min_c \sum_i |x_i - c|$. the solution is the median of 2, 4, 9 (you can also do trial and error), i.e. the 4th pole.

3. (Probability) In your 10-day vacation to Dakota, you kept the following count of days:

rainy	warm	bighorn (saw sheep)	days
N	N	N	1
N	N	Y	0
N	Y	N	0
N	Y	Y	4
Y	N	N	1
Y	N	Y	1
Y	Y	N	1
Y	Y	Y	2

   Use maximum likelihood estimate (no smoothing), estimate the probability $P(bighorn = Y \mid rainy = N, warm = N)$.
   A: $P(bighorn = Y \mid rainy = N, warm = N) = c(NNY)/(c(NNN) + c(NNY)) = 0$.

4. (MDP) Consider state space $S = \{s_1, s_2\}$ and action space $A = \{left, right\}$:

s_1	s_2

   In s_1 the action "right" sends the agent to s_2 and collects reward $r = 1$. In s_2 the action "left" sends the agent to s_1 but with zero reward. All other state-action pairs stay in that state with zero reward. With discounting factor γ, what is the value $v(s_2)$ under the optimal policy?
   A: The optimal policy is $\pi(s_2) = left, \pi(s_1) = right$.

   $v(s_2) = 0 + \gamma \cdot 1 + \gamma^2 \cdot 0 + \gamma^3 \cdot 1 + ... = \gamma + \gamma^3 + \gamma^5 + ...$(OK to stop here) $= \gamma/(1 - \gamma^2)$.

5. (Q-learning) A robot initializes Q-learning by setting $q(s, a) = 0$ for all state s and action a. It has a learning rate α, and discounting factor γ. The robot senses that it is in state s_{105} and decides to performs action a_{540}. For this action, the robot receives

reward 100 and arrives at state s_{7331}. What value is $q(s_{105}, a_{540})$ after this one step of Q-learning?
A: $q(s_{105}, a_{540}) = \alpha 100$. All other things zero out.

6. (Game) Consider a variant of the II-nim game. There are two piles, each pile has two sticks. A player can take one stick from a single pile; or take two sticks, one from each pile (when available). The player who takes the last stick wins. Let the game value be 1 if the first player wins. What is the game theoretical value of this game?
A: -1. Draw the game tree. (It is true that we didn't specify the value if the 2nd player wins. However, 0 is never a reasonable answer. It has to be something negative.)

7. (Stepsize) We use gradient descent to find the minimum of the function $f(x) = x^2$ with step size $\eta > 0$. If we start from the point $x_0 = 3$, how small should η be so we make progress in each iteration?
A: $\eta < 1$. You can get this by requiring $f(x_1) < f(x_0)$, where $x_1 = x_0 - \eta f'(x) = x_0 - 2\eta x_0$.

8. (Resolution) Given knowledge base

   (a) $P \Leftrightarrow Q$

   (b) P

   use resolution to prove query Q.

```
KB CNF:
1 ~P v Q
2 ~Q v P
3 P
negated query CNF:
4 ~Q
1+4=5 ~P
5+3=6 null
```

9. (FOL) Which one is the translation of "Frodo has exactly one ring"?
(A) $\exists x, y \ HasRing(Frodo, x) \wedge HasRing(Frodo, y) \wedge x = y$
(B) $\forall x \ HasRing(Frodo, x) \Rightarrow \exists y (HasRing(Frodo, y) \wedge x = y)$
(C) $\exists x \ HasRing(Frodo, x) \Rightarrow \forall y (HasRing(Frodo, y) \wedge x = y)$
(D) $\exists x \ HasRing(Frodo, x) \wedge \forall y (HasRing(Frodo, y) \Rightarrow x = y)$
(E) none of the above
A: D

10. (PCA) You performed PCA in R^3. It turns out that the first principal component is $u_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$, and the second principal component is $u_2 = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$. One of your original data points was $x = (1, 2, 3)$. What will its new representation be?
A: $(u_1^\top x, u_2^\top x) = (\frac{3}{\sqrt{2}}, \frac{1}{\sqrt{2}})$

11. (Linear regression) Long time ago, a primate researcher gave you a data set to predict the label y (monkey daily diet weight) from a number of features $x_1, \ldots, x_d$. You built a linear regression model for him:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d.$$

Yesterday, the monkey researcher realized that his RA mixed things up: x_d was actually the label, while y was the d-th feature! Alas, neither him nor you have the data set anymore, and the RA was long gone to start up a monkey intelligence company and does not respond to emails. All you have are the coefficients $\beta_0 \ldots \beta_d$, all non-zero. How do you fix the linear regression model? (One line math)
A:

$$(y - \beta_0 - \beta_1 x_1 - \ldots - \beta_{d-1} x_{d-1})/\beta_d = x_d.$$

12. (A*) Consider A* search on the following state space, with initial state A and goal state G, and one can move left, right, up, or down one step at a time (but no wrapping around). The cost g is the number of moves taken, and the heuristic h is the Manhattan distance to G.

A	B
G	C

At the moment that A* declares success, which states remain in OPEN?

```
OPEN:
(A 0+1)
(B 1+2, G 1+0)
(B 1+2) <--
```

13. (Search) Let the search space be natural numbers. Each state n has three successors: $3n, 3n+1, 3n+2$. How many shortest paths (i.e. the sequence of states) are there from the initial state 1 to the goal state 540?
A: Zero. The path to 540 is 2, 6, 20, 60, 180, 540, which does not start from 1.

14. (Iterative Deepening) Consider a search graph which is a tree, and each internal node has b children. The only goal node is at depth $d = 3$ (root is depth 1). How many total goal-checks will be preformed by iterative deepening in the luckiest case (i.e. the smallest number of goal-checks)? If a node is checked multiple times you should count that multiple times.
A: Luckiest means the goal is the first node checked at depth d. consider $d = 2$, the goal-checks are: root; root, goal. If $d = 3$: root; root, child1, ..., child-b; root, child1, goal. That's $1 + (1 + b) + (1 + 1 + 1) = b + 5$ goal checks. We will also accept $b + 4$ because the slides did not count the root by itself as the first round.

15. (Gradient descent) Let $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. We want to minimize the objective function $f(\mathbf{x}) = \sum_{i=1}^{d} i x_i = x_1 + 2x_2 + \ldots + d x_d$ using gradient descent. Let the stepsize $\eta = 0.1$. If we start at the all-zero vector $\mathbf{x}^{(0)} = (0, \ldots, 0)$, what is the next vector $\mathbf{x}^{(1)}$ produced by gradient descent?
A: $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \eta \nabla f(\mathbf{x})$, and $\nabla f(\mathbf{x}) = (1, 2, \ldots, d)$. So $(-0.1, -0.2, \ldots, -0.1 \cdot d)$.

16. (Conjunctive normal form) Write down the CNF for $(P \Rightarrow Q) \Rightarrow (Q \Rightarrow R)$.

$$(\neg P \vee Q) \Rightarrow (\neg Q \vee R)$$
$$\neg(\neg P \vee Q) \vee (\neg Q \vee R)$$
$$(P \wedge \neg Q) \vee (\neg Q \vee R)$$

So the answer is

$$(P \vee \neg Q \vee R) \wedge (\neg Q \vee R)$$

17. (Clustering) There are five points in one-dimensional space: $a = 0, b = 1, c = 3, d = 7, e = 9$. Perform Hierarchical Agglomerative Clustering with complete linkage. Complete the resulting clustering tree diagram (i.e., the dendrogram).

$a = 0 \qquad b = 1 \qquad c = 3 \qquad d = 7 \qquad e = 9$

```
(ab) c d e
(ab) c (de)
(ab) c: 3, c (de): 6
((ab) c) (de)
(((ab) c) (de))
```

18. (kNN) Let a data set consists of $n+1$ points in $\mathbb{R}$, specifically $x_1 = 1, \ldots, x_n = n$, and $x_{n+1} > n$. What is the threshold on the value of x_{n+1}, above which x_n is among x_{n+1}'s 3-nearest neighbors, but x_{n+1} is NOT among x_n's 3-nearest neighbors?
A: $x_{n+1} > n+3$. Note the 3NN doesn't include a point itself. However, if you assumed that 3NN includes the point itself, this would be $x_{n+1} > n+2$.

19. (Neural Network) Let the input $x \in R$. Thus the input layer has a single x input. The network has 5 hidden layers. Each hidden layer has 10 units. The output layer has a single unit and outputs $y \in R$. Between layers the network is fully connected. All units in the network have a bias 1 input. All units are LINEAR units, namely the activation function is the identity function: $v = f(u) = u$, while u is a linear combination of all inputs to that unit (including the bias 1). What kind of functions can this network compute?
(a) $y = ax + b$ (b) $y = ax^2 + bx + c$ (c) $y = \frac{1}{1+e^{-\theta x}}$ (d) $y = \max(0, x)$
A: (a). All linear functions $y = ax + b$. Composition of linear functions is linear.

20. (Probability) There are two biased coins in my wallet: coin A has $P(Heads) = 0.1$, coin B has $P(Heads) = 0.8$. I took out one coin at random (with equal probability choosing A or B) and flipped it once: the outcome was Heads. What is the probability that the coin was A?
A: $P(A \mid H) = P(H \mid A)P(A)/(P(H \mid A)P(A) + P(H \mid B)P(B)) = 0.1 * 0.5/(0.1 * 0.5 + 0.8 * 0.5) = 1/9$.

21. (Accuracy) You trained a binary classifier on a training set, and tested its performance on a separate test set. Your classifier was wrong on every test item! You are very happy: why? (One sentence)
A: Negate the predictions of your classifier, you then have a very good classifier.

22. (k-means) Consider four points in $\mathbb{R}^2$: $x_1 = (0, 0)$, $x_2 = (0, 1)$, $x_3 = (1, 0)$, $x_4 = (1, 1)$. Let there be two initial cluster centers $c_1 = (0, 0), c_2 = (3, 3)$. Use Euclidean distance. Tell us what happens to the cluster centers after one iteration of k-means. (Two sentences at most)
A: $c_1 = (0.5, 0.5)$ which grabs all points; c_2 undefined since it loses all points. If you say you will implement it so that a cluster stays the same if it loses all points and thus you get (3,3), we will accept that, too.

23. (ReLU) Consider a rectified linear unit with input $x \in \mathbb{R}$ and a bias term. The output can be written as $y = \max(0, w_0 + w_1 x)$. Write down the input value x that produces a specific output $y > 0$.
A: since $y > 0, y = w_0 + w_1 * x$. Thus $x = (y - w_0)/w_1$. However, we will also accept the interpretation that we asked for the range $y > 0$. In that case you must show all four branches:
$$\begin{cases} x > -w_0/w_1, & w_1 > 0 \\ x < -w_0/w_1, & w_1 < 0 \\ x \in \mathbb{R}, & w_1 = 0, w_0 > 0 \\ \emptyset, & w_1 = 0, w_0 \leq 0 \end{cases}$$

24. (Search) Consider $n+1$ states. S_1 is the initial state, S_n is the goal state. S_0 is a dead-end state with no successors. For each state S_i, $i = 1, \ldots, n-1$, it has two successors: S_{i+1} and S_0. S_n also has no successors. There is no cycle check nor CLOSED list. How many goal-checks will be performed by the two search algorithms: breadth first search, and iterative deepening depth first search, respectively? (Assume everything being equal, state with small index is checked first. If a state is goal-checked multiple times, count it multiple times.)
A:
BFS: s1, s0,s2, s0, s3, ..., s0, sn. $(n - 1) * 2 + 1 = 2n - 1$
ID: s1; s1,s0,s2; ...; (s1, s0,s2, s0, s3, ..., s0, sn). $1+3+5+..+(2n-1) = \sum_{i=1}^{n}(2i-1) = n^2$.

25. (A*) Given two admissible heuristic functions h_1, h_2, what range of α guarantees that

$h_1 - \alpha(h_1 - h2)$ is admissible?
A: $\alpha \in [0, 1]$, this is the convex combination (weighted average) of the two heuristics.