CSYE 7245

Big-Data Systems and Intelligence Analytics

Assignment 3

Q1 (5 Points)

Give a brief definition for the following:

1. Linearity of expectation

Linearity of expectation is the property that the expected value of the sum of random variables is equal to the sum of their individual expected values, regardless of whether they are independent.

1. Z-score

Z-score is the number of standard deviations from the mean a data point is. But more technically it’s a measure of how many standard deviations below or above the population mean a raw score is. It is also known as a standard score and it can be placed on a normal distribution curve.

1. Chernoff bound

The Chernoff bound is like a genericized trademark in that it refers not to a particular inequality, but rather a technique for obtaining exponentially decreasing bounds on tail probabilities.

1. Monte Carlo algorithm

The Monte Carlo algorithm uses a finite resource path to generate the above-mentioned “fuzzy” results with a margin of error. Monte Carlo algorithms often rely on repeated random sampling – they get general random numbers, and look for probability in order to provide results.

1. Bias-variance tradeoff

The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. The error due to variance is taken as the variability of a model prediction for a given data point. If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it’s going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data. This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

Q2 (5 Points)

Arrange the following functions in increasing order of asymptotic growth:

* n2
* 0.33n
* 5n5
* n2 √n
* 5n
* log n
* √n

0.33n < log(n) < √n (n1/2) < n2 < n2√n (n5/2) < 5n < 5n5

Q3 (5 Points)

Master Theorem: For the following recurrence, give an expression for the runtime T(n) if the recurrence can be solved with the Master Theorem. Otherwise, indicate that the Master Theorem does not apply.

T(n) <= aT(n/b)+f(n)

1. T(n) = 4T (n/2)+ n a = 4, b = 2

f(n) = n = O(n^(logba)-e) = O(n2-1) (case 1)

T(n) = Θ(n2)

1. T(n) = 4T (n/2)+ n2 a = 4, b = 2

f(n) = n2 = Θ(n^logba \* logkn) = Θ(n2 \* log0n)

T(n)= Θ(n^logba \* logk+1n)= Θ (n2logn) (case 2)

1. T(n) = 2T (n/2)+ n2 log n a = 2, b = 2

f(n) = n2 log n != Θ(n^logba \* logkn)

f(n) = n2 log n != Θ(n^logba -e)

Not apply

Q4 (5 Points)

Stephen Curry hit 77 three-point shots in a row in practice. If his probability of hitting an unguarded three-point shot is 90%, what is the likelihood of Stephen Curry making at least 9 out of 10 three-point shots?

Likelihood = 0.99 = 0.38742

Q5 (5 Points)

A booth at the fair has 200 balloons, 5 of which contain $10 and 1 of which contains $20. The rest contain only air. If it costs $1 to randomly break a balloon, what is the expected return of an individual making such an attempt?

Expectation = 1/200 \* 20 + 5/200 \* 10 - 1 = -0.65

Q6 (5 Points)

Sort the list of integers below using Merge sort. Show your work. Write a recurrence relation for Merge sort.

(22, 13, 26, 1, 12, 27, 33, 15)

Use the Master Theorem to prove the complexity of your recurrence.

Recurrence relation: T(n) = 2 T(n/2) + f(n)

T(n) = 2T (n/2)+ n a = 2, b = 2

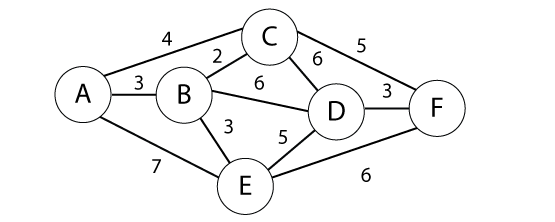
f(n) = nlogn = Θ(n^logba \* logkn) = Θ(n \* log1n) (case 2)

T(n) = Θ(nlogn)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Original | 22 | 13 | 26 | 1 | 12 | 27 | 33 | 15 |  |  |  |  |  |  |  |
| Divide in 2 | 22 | 13 | 26 | 1 |  | 12 | 27 | 33 | 15 |  |  |  |  |  |  |
| Divide in 4 | 22 | 13 |  | 26 | 1 |  | 12 | 27 |  | 33 | 15 |  |  |  |  |
| Divide in 8 | 22 |  | 13 |  | 26 |  | 1 |  | 12 |  | 27 |  | 33 |  | 15 |
| Merge 1 | 13 | 22 |  |  | 1 | 26 |  |  | 12 | 27 |  |  | 15 | 33 |  |
| Merge 2 | 1 | 13 | 22 | 26 |  |  |  |  | 12 | 15 | 27 | 33 |  |  |  |
| Merge 3 | 1 | 12 | 13 | 15 | 22 | 26 | 27 | 33 |  |  |  |  |  |  |  |

Q7 (5 Points)

Find shortest path from A to F in the graph below using Dijkstra's algorithm. *Show your steps.*

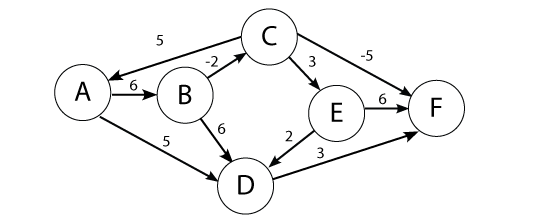


|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F |
| 1 A {A} |  | (3, A) \*\* | (4, A) | INF | (7, A) | INF |
| 2 B {A, B} |  |  | (4, A) \*\* | (9, B) | (6, B) | INF |
| 3 C {A, B, C} |  |  |  | (9, B) | (6, B) \*\* | (9, C) |
| 4 E {A, B, C, E} |  |  |  | (9, B) \*\* |  | (9, C) |
| 5 D {A, B, C, E, D} |  |  |  |  |  | (9, C) \*\* |
| 6 F {A, B, C, E, D, F} |  | (3, A) | (4, A) | (9, B) | (6, B) | (9, C) |

Shortest path: A -> C -> F

Length = 9

Q8 (5 Points)



Use the Bellman-Ford algorithm to find the shortest path from node A to F in the weighted directed graph above. *Show your work.*

Edges: AB, AD, CA, BC, BD, CE, ED, CF, DF, EF

Shortest path: A->B->C->F at cost -1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F |
| 0 | 0 | INF | INF | INF | INF | INF |
| 1 | 0 | 6 | INF | 5 | INF | INF |
| 2 | 0 | 6 | 4 | 5 | INF | 8 |
| 3 | 0 | 6 | 4 | 5 | 7 | -1 |
| 4 | 0 | 6 | 4 | 5 | 7 | -1 |

Q9 (5 Points)

In a room of 23 people, what is the probability that someone has the same birthday as you?

Probability = 1 - 365\*364\* …… \* 342 \* 343 / 36523 = 1 - 0.4927 = 0.5073

Q10 (5 Points)

Suppose S is a very large set of real numbers, and you would like to estimate its median value by sampling. (It is too expensive to sort S to calculate the exact median). You may assume all the numbers in S are distinct. Let n =|S|; a number x is an *ε-approximate median* if at least (0.5 – ε)n numbers in S are less than x and (0.5 – ε)n numbers in S are greater than x.

Suppose you sample a subset uniformly at random (with replacement) and use that to estimate your median. Use a Chernoff-bound to calculate your confidence in your approximate median estimate. You can choose your confidence level (typical is 90%, 95% or 99% confidence) the distribution of the numbers and your sample size.

For δ = 0.1 and μ = x

The Chernoff bound (above mean):  
Pr[X>(1+ δ) μ] < [eδ /(1+ δ) (1+ δ)] μ   
Pr[X>(1+ 0.1) x] < [e 0.1 /(1+ 0.1) (1+ 0.1)] x

The Chernoff bound (below mean)  
Pr[X<(1- δ) μ] < [e-δ^2 μ/2 /(1+ δ) (1+ δ)] μ   
Pr[X<(1- 0.1) x] < [e-0.1^2 x/2 /(1+ 0.1) (1+ 0.1)] x

Q11 (5 Points)

Two linear regression models return t-statistics of 1 and 19 respectively. What is the null hypothesis in this case. Which t-statistic provides more evidence to reject the null hypothesis.

y = β0 + β1x1 + β2x2 + · · · + βkxk

In this case, the null hypothesis is: β1 = β2 = · · · = βk =0

T-score = 1, degree of freedom = 1 🡪 Probability = 0.25

T-score = 19, degree of freedom = 1 🡪 Probability = 0.0167

The second model t-statistic provides more evidence to reject the null hypothesis.

Q12 (5 Points)

Given the weights and values of the four items in the table below, select a subset of items with the maximum combined value that will fit in a knapsack with a weight limit, *W,* of 6. Use dynamic programming. *Show your work.*

|  |  |  |
| --- | --- | --- |
| Item i | Value vi | Weight wi |
| 1  2  3  4 | 3  2  4  4 | 4  3  2  3 |

Capacity of knapsack W=6

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Items/weight | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 3 | 3 | 3 |
| 2 | 0 | 0 | 0 | 2 | 3 | 3 | 3 |
| 3 | 0 | 0 | 4 | 4 | 4 | 6 | 7 |
| 4 | 0 | 0 | 4 | 4 | 4 | 8 | 8 |

Max Value = 8

Max Items = [3, 4]

Q13 (40 Points) Search in Pacman

For this question, you’ll be implementing search in Pacman (<http://ai.berkeley.edu/search.html> ) from The Pac-Man projects were developed for UC Berkeley's introductory artificial intelligence course, CS 188 (http://ai.berkeley.edu/project\_overview.html).

Your Pacman agent will find paths through his maze world, both to reach a particular location and to collect food efficiently. You will build general search algorithms and apply them to Pacman scenarios.

As in Project 0, this project includes an autograder for you to grade your answers on your machine. This can be run with the command:

python autograder.py

See the autograder tutorial in Project 0 for more information about using the autograder.

The code for this project consists of several Python files, some of which you will need to read and understand in order to complete the assignment, and some of which you can ignore. You can download all the code and supporting files as a [zip archive](https://s3-us-west-2.amazonaws.com/cs188websitecontent/projects/release/search/v1/001/search.zip).

Q12-1: Depth First Search

Q12-2: Breadth First Search

Q12-3: Uniform Cost Search

Q12-4: A\* Search

Q12-5: Corners Problem: Representation

Q12-6: Corners Problem: Heuristic

Q12-7: Eating All The Dots: Heuristic

Q12-8: Suboptimal Search

### Question 1 (5 points): Finding a Fixed Food Dot using Depth First Search

In searchAgents.py, you'll find a fully implemented SearchAgent, which plans out a path through Pacman's world and then executes that path step-by-step. The search algorithms for formulating a plan are not implemented -- that's your job. As you work through the following questions, you might find it useful to refer to the object glossary (the second to last tab in the navigation bar above).

First, test that the SearchAgent is working correctly by running:

python pacman.py -l tinyMaze -p SearchAgent -a fn=tinyMazeSearch

The command above tells the SearchAgent to use tinyMazeSearch as its search algorithm, which is implemented in search.py. Pacman should navigate the maze successfully.

Now it's time to write full-fledged generic search functions to help Pacman plan routes! Pseudocode for the search algorithms you'll write can be found in the lecture slides. Remember that a search node must contain not only a state but also the information necessary to reconstruct the path (plan) which gets to that state.

**Important note:** All of your search functions need to return a list of actions that will lead the agent from the start to the goal. These actions all have to be legal moves (valid directions, no moving through walls).

**Important note:** Make sure to **use** the Stack, Queue and PriorityQueue data structures provided to you in util.py! These data structure implementations have particular properties which are required for compatibility with the autograder.

Hint: Each algorithm is very similar. Algorithms for DFS, BFS, UCS, and A\* differ only in the details of how the fringe is managed. So, concentrate on getting DFS right and the rest should be relatively straightforward. Indeed, one possible implementation requires only a single generic search method which is configured with an algorithm-specific queuing strategy. (Your implementation need not be of this form to receive full credit).

Implement the depth-first search (DFS) algorithm in the depthFirstSearch function in search.py. To make your algorithm complete, write the graph search version of DFS, which avoids expanding any already visited states.

Your code should quickly find a solution for:

python pacman.py -l tinyMaze -p SearchAgent

python pacman.py -l mediumMaze -p SearchAgent

python pacman.py -l bigMaze -z .5 -p SearchAgent

The Pacman board will show an overlay of the states explored, and the order in which they were explored (brighter red means earlier exploration). Is the exploration order what you would have expected? Does Pacman actually go to all the explored squares on his way to the goal?

Hint: If you use a Stack as your data structure, the solution found by your DFS algorithm for mediumMaze should have a length of 130 (provided you push successors onto the fringe in the order provided by getSuccessors; you might get 246 if you push them in the reverse order). Is this a least cost solution? If not, think about what depth-first search is doing wrong.

### Question 2 (5 points): Breadth First Search

Implement the breadth-first search (BFS) algorithm in the breadthFirstSearch function in search.py. Again, write a graph search algorithm that avoids expanding any already visited states. Test your code the same way you did for depth-first search.

python pacman.py -l mediumMaze -p SearchAgent -a fn=bfs

python pacman.py -l bigMaze -p SearchAgent -a fn=bfs -z .5

Does BFS find a least cost solution? If not, check your implementation.

Hint: If Pacman moves too slowly for you, try the option --frameTime 0.

Note: If you've written your search code generically, your code should work equally well for the eight-puzzle search problem without any changes.

python eightpuzzle.py

### Question 3 (5 points): Varying the Cost Function

While BFS will find a fewest-actions path to the goal, we might want to find paths that are "best" in other senses. Consider mediumDottedMaze and mediumScaryMaze.

By changing the cost function, we can encourage Pacman to find different paths. For example, we can charge more for dangerous steps in ghost-ridden areas or less for steps in food-rich areas, and a rational Pacman agent should adjust its behavior in response.

Implement the uniform-cost graph search algorithm in the uniformCostSearch function in search.py. We encourage you to look through util.py for some data structures that may be useful in your implementation. You should now observe successful behavior in all three of the following layouts, where the agents below are all UCS agents that differ only in the cost function they use (the agents and cost functions are written for you):

python pacman.py -l mediumMaze -p SearchAgent -a fn=ucs

python pacman.py -l mediumDottedMaze -p StayEastSearchAgent

python pacman.py -l mediumScaryMaze -p StayWestSearchAgent

Note: You should get very low and very high path costs for the StayEastSearchAgent and StayWestSearchAgent respectively, due to their exponential cost functions (see searchAgents.py for details).

### Question 4 (5 points): A\* search

Implement A\* graph search in the empty function aStarSearch in search.py. A\* takes a heuristic function as an argument. Heuristics take two arguments: a state in the search problem (the main argument), and the problem itself (for reference information). The nullHeuristic heuristic function in search.py is a trivial example.

You can test your A\* implementation on the original problem of finding a path through a maze to a fixed position using the Manhattan distance heuristic (implemented already as manhattanHeuristic in searchAgents.py).

python pacman.py -l bigMaze -z .5 -p SearchAgent -a fn=astar,heuristic=manhattanHeuristic

You should see that A\* finds the optimal solution slightly faster than uniform cost search (about 549 vs. 620 search nodes expanded in our implementation, but ties in priority may make your numbers differ slightly). What happens on openMaze for the various search strategies?

### Question 5 (5 points): Finding All the Corners

The real power of A\* will only be apparent with a more challenging search problem. Now, it's time to formulate a new problem and design a heuristic for it.

In corner mazes, there are four dots, one in each corner. Our new search problem is to find the shortest path through the maze that touches all four corners (whether the maze actually has food there or not). Note that for some mazes like tinyCorners, the shortest path does not always go to the closest food first! Hint: the shortest path through tinyCorners takes 28 steps.

*Note: Make sure to complete Question 2 before working on Question 5, because Question 5 builds upon your answer for Question 2.*

Implement the CornersProblem search problem in searchAgents.py. You will need to choose a state representation that encodes all the information necessary to detect whether all four corners have been reached. Now, your search agent should solve:

python pacman.py -l tinyCorners -p SearchAgent -a fn=bfs,prob=CornersProblem

python pacman.py -l mediumCorners -p SearchAgent -a fn=bfs,prob=CornersProblem

To receive full credit, you need to define an abstract state representation that does not encode irrelevant information (like the position of ghosts, where extra food is, etc.). In particular, do not use a Pacman GameState as a search state. Your code will be very, very slow if you do (and also wrong).

Hint: The only parts of the game state you need to reference in your implementation are the starting Pacman position and the location of the four corners.

Our implementation of breadthFirstSearch expands just under 2000 search nodes on mediumCorners. However, heuristics (used with A\* search) can reduce the amount of searching required.

### Question 6 (5 points): Corners Problem: Heuristic

*Note: Make sure to complete Question 4 before working on Question 6, because Question 6 builds upon your answer for Question 4.*

Implement a non-trivial, consistent heuristic for the CornersProblem in cornersHeuristic.

python pacman.py -l mediumCorners -p AStarCornersAgent -z 0.5

Note: AStarCornersAgent is a shortcut for

-p SearchAgent -a fn=aStarSearch,prob=CornersProblem,heuristic=cornersHeuristic.

**Admissibility vs. Consistency:** Remember, heuristics are just functions that take search states and return numbers that estimate the cost to a nearest goal. More effective heuristics will return values closer to the actual goal costs. To be admissible, the heuristic values must be lower bounds on the actual shortest path cost to the nearest goal (and non-negative). To be consistent, it must additionally hold that if an action has cost c, then taking that action can only cause a drop in heuristic of at most c.

Remember that admissibility isn't enough to guarantee correctness in graph search -- you need the stronger condition of consistency. However, admissible heuristics are usually also consistent, especially if they are derived from problem relaxations. Therefore it is usually easiest to start out by brainstorming admissible heuristics. Once you have an admissible heuristic that works well, you can check whether it is indeed consistent, too. The only way to guarantee consistency is with a proof. However, inconsistency can often be detected by verifying that for each node you expand, its successor nodes are equal or higher in in f-value. Moreover, if UCS and A\* ever return paths of different lengths, your heuristic is inconsistent. This stuff is tricky!

**Non-Trivial Heuristics:** The trivial heuristics are the ones that return zero everywhere (UCS) and the heuristic which computes the true completion cost. The former won't save you any time, while the latter will timeout the autograder. You want a heuristic which reduces total compute time, though for this assignment the autograder will only check node counts (aside from enforcing a reasonable time limit).

**Grading:** Your heuristic must be a non-trivial non-negative consistent heuristic to receive any points. Make sure that your heuristic returns 0 at every goal state and never returns a negative value. Depending on how few nodes your heuristic expands, you'll be graded:

|  |  |
| --- | --- |
| **Number of nodes expanded** | **Grade** |
| more than 2000 | 0/5 |
| at most 2000 | 1/5 |
| at most 1600 | 3/5 |
| at most 1200 | 5/5 |

Remember: If your heuristic is inconsistent, you will receive no credit, so be careful!

### Question 7 (4 points): Eating All The Dots

Now we'll solve a hard search problem: eating all the Pacman food in as few steps as possible. For this, we'll need a new search problem definition which formalizes the food-clearing problem: FoodSearchProblem in searchAgents.py (implemented for you). A solution is defined to be a path that collects all of the food in the Pacman world. For the present project, solutions do not take into account any ghosts or power pellets; solutions only depend on the placement of walls, regular food and Pacman. (Of course ghosts can ruin the execution of a solution! We'll get to that in the next project.) If you have written your general search methods correctly, A\* with a null heuristic (equivalent to uniform-cost search) should quickly find an optimal solution to testSearch with no code change on your part (total cost of 7).

python pacman.py -l testSearch -p AStarFoodSearchAgent

Note: AStarFoodSearchAgent is a shortcut for -p SearchAgent -a fn=astar,prob=FoodSearchProblem,heuristic=foodHeuristic.

You should find that UCS starts to slow down even for the seemingly simple tinySearch. As a reference, our implementation takes 2.5 seconds to find a path of length 27 after expanding 5057 search nodes.

*Note: Make sure to complete Question 4 before working on Question 7, because Question 7 builds upon your answer for Question 4.*

Fill in foodHeuristic in searchAgents.py with a consistent heuristic for the FoodSearchProblem. Try your agent on the trickySearchboard:

python pacman.py -l trickySearch -p AStarFoodSearchAgent

Our UCS agent finds the optimal solution in about 13 seconds, exploring over 16,000 nodes.

Any non-trivial non-negative consistent heuristic will receive 1 point. Make sure that your heuristic returns 0 at every goal state and never returns a negative value. Depending on how few nodes your heuristic expands, you'll get additional points:

|  |  |
| --- | --- |
| **Number of nodes expanded** | **Grade** |
| more than 15000 | 1/4 |
| at most 15000 | 2/4 |
| at most 12000 | 3/4 |
| at most 9000 | 4/4 (full credit; medium) |
| at most 7000 | 5/4 (optional extra credit; hard) |

Remember: If your heuristic is inconsistent, you will receive no credit, so be careful! Can you solve mediumSearch in a short time? If so, we're either very, very impressed, or your heuristic is inconsistent.

### Question 8 (5 points): Suboptimal Search

Sometimes, even with A\* and a good heuristic, finding the optimal path through all the dots is hard. In these cases, we'd still like to find a reasonably good path, quickly. In this section, you'll write an agent that always greedily eats the closest dot. ClosestDotSearchAgent is implemented for you in searchAgents.py, but it's missing a key function that finds a path to the closest dot.

Implement the function findPathToClosestDot in searchAgents.py. Our agent solves this maze (suboptimally!) in under a second with a path cost of 350:

python pacman.py -l bigSearch -p ClosestDotSearchAgent -z .5

Hint: The quickest way to complete findPathToClosestDot is to fill in the AnyFoodSearchProblem, which is missing its goal test. Then, solve that problem with an appropriate search function. The solution should be very short!

Your ClosestDotSearchAgent won't always find the shortest possible path through the maze. Make sure you understand why and try to come up with a small example where repeatedly going to the closest dot does not result in finding the shortest path for eating all the dots.