Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Artificial Intelligence (First Year)

Main Summer Examinations 2019

Time allowed: 2:00

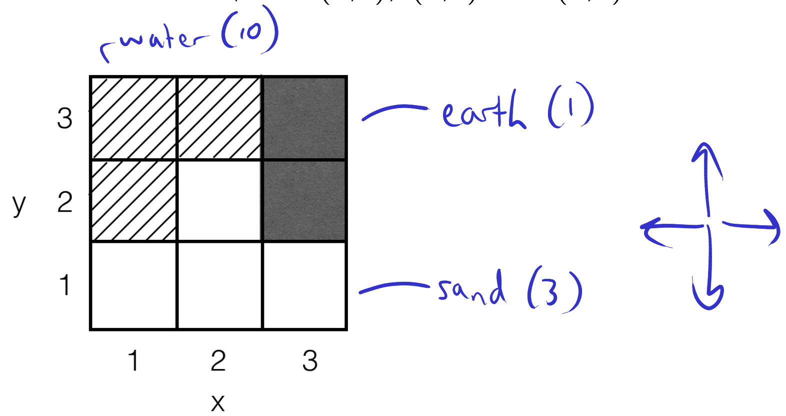
[Answer all questions]

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 80, which will be rescaled to a mark out of 100.

Question 1 Search and Optimisation

Assume that you are developing an algorithm to find the lowest cost path to move an army from a starting to a goal position in a strategy game. The field is organised as a grid, where the positions whose coordinates (x, y) are (3,2) and (3,3) have earth terrain; (1,1), (2,1), (3,1) and (2,2) have sand terrain; and (1,2), (1,3) and (2,3) have water terrain:



Each move can take the army one position up, down, left or right on the grid, so long as this does not move the army outside the grid. Independent of the type of terrain of the current position of the army, the cost of moving to an earth, sand and water position is 1, 3 and 10, respectively. For example, the cost of moving left from (3,2) is 3.

(a) Your army is in position (3,1) and you wish to reach the goal position (1,3). Assume that you have decided to use a breadth-first search algorithm to solve this problem, respecting the following rules:

- A state in the state space graph is identified by the (x, y) coordinates of the current position of the army, meaning that there are 9 possible states.
- Do not place children in the frontier if their corresponding state is already in the frontier or list of visited nodes.
- Stop when you place in the frontier a node which contains the goal state.
- When deciding which node to visit, if there is a draw, choose to visit the node with the smallest x-coordinate first. If this still results in a draw, choose to visit the node with the smallest y-coordinate first. For example, if there is a draw between nodes whose states are (3,1) and (2,2), visit (2,2) first. If there is a draw between (3,1) and (3,2), visit (3,1) first.

Question 1 continued over the page

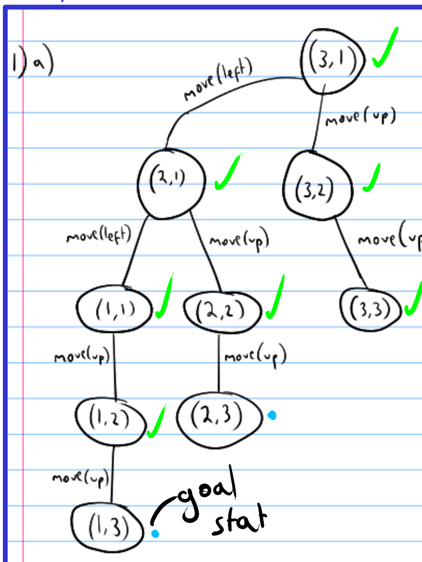
Write down the following information:

- Search tree produced by breadth-first search. Indicate which nodes are visited nodes and which nodes are in the frontier when the algorithm terminates.
- Sequence of nodes *visited* by breadth-first search. Note: you can identify a node through its state.
- Sequence of states that compose the path retrieved as a solution by breadth-first search.

[8 marks]

Shay

1) a)



when the algorithm terminates:

visited : (3,1), (2,1), (3,2), (1,1), (2,2), (3,3), (1,2)

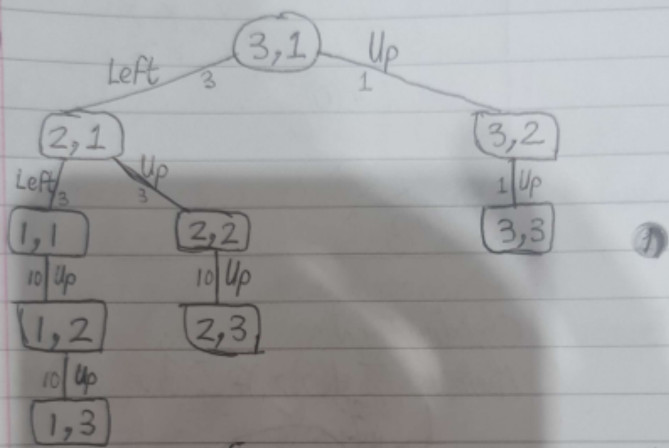
frontier : (2,3), (1,3)

Solution : move(left), move(left), move(up), move(up)

Path to solution: (3,1), (2,1), (1,1), (1,2), (1,3)

Adwit

1a)



- Visited Nodes = {(3,1), (2,1), (3,2), (1,1), (2,2), (3,3), (1,2)}
- Frontier at termination: {(1,3), (2,3)}
- Sequence of nodes visited = [(3,1), (2,1), (3,2), (1,1), (2,2), (3,3), (1,2)]
- Path of Solution: [(3,1), (2,1), (1,1), (1,2), (1,3)]

Cost: 26

(b) Is breadth-first search a good algorithm for this problem? Justify your answer.

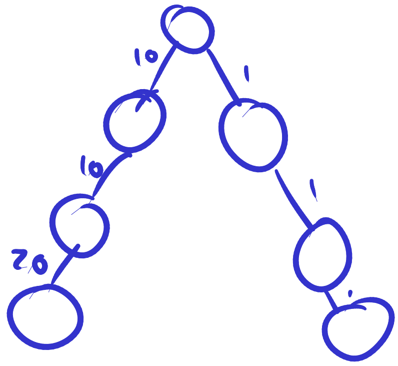
(DFS better in this case

[6 marks]

Nisar

nisar Today at 13:40

No as the algorithm it's not optimal. For BFS to be optimal the path cost should be a nondrecreasing function of the depth of the node, so we move deeper into the tree (or graph), the cost of reaching a node should not decrease. In this case, as each action has a different cost, this means BFS may not be the most optimal in this case.



Shay

b) Breadth-First Search may not be a good algorithm as this problem scales up, as we're trying to expand every possible route, even if it has a high cost. We may want to use A* or Dijkstra, which are specialized for pathfinding. BFS does not consider the different terrain costs.

Leonardo's slides

- **Optimality**: BFS is optimal if the path cost is a nondecreasing function of the depth of the node (e.g., all actions have the same cost)

(c) Consider that you wish to use Hill Climbing to solve this problem. For that, you need to provide a problem formulation.

- Specify the design variable of your problem formulation.
- Explain your design variable by giving two examples of candidate solutions. If infeasible solutions are possible, give one example of feasible, and one example of infeasible solution, explaining the reason for them to be feasible / infeasible. If no infeasible solutions exist given your problem formulation, give two examples of feasible solutions, and explain why no infeasible solutions exist.

[6 marks]

Shay

c) Our design variable could be a sequence of actions, e.g. (left, left, up, up), stored as a vector.

One feasible candidate solution would be: (left, left, up, up). This is feasible since if we follow these actions from our initial state (3,1) we will reach our goal node (1,3).

One infeasible candidate solution would be: (left, up, right), as we would not reach our goal state in this case.

Adwit

1c) The design variable would be a vector (x) where x contains the nodes visited on the path to the goal (not including initial state and goal).

- Feasible Solution: $x = [(2,1),(1,1),(1,2)]$
  This is feasible because the start of the path is adjacent to the initial square on the grid meaning its a valid move. Similarly, every x_i and x_{i+1} squares are adjacent and the final node is adjacent to the goal.
- Infeasible Solution: $x = [(3,3),(3,2),(2,5)]$
  This is an infeasible solution for 4 reasons:
  - (3,3) is not adjacent to initial state
  - (3,2) is not adjacent to (2,5)
  - (2,5) is not a state (not in grid)
  - (2,5) is not adjacent to goal state

meaning there's many constraints violated due to invalid moves.

Question 2 k-Nearest Neighbours and Naïve Bayes

tweet

(a) Consider the problem of predicting the best hashtag to be associated to a tweet. To solve this problem, you have access to incoming tweets that can be used as training examples. Each hashtagged tweet is a training example. Every second, on average, around 6,000 tweets are tweeted on Twitter, and most of them use hashtags. Each tweet is described by a set of input attributes and one output attribute. There is one categorical input attribute for each possible word that can appear in a tweet. Each input attribute can assume values *true* or *false*, representing whether or not the corresponding word appears in the tweet. The output attribute is a categorical value corresponding to the first hashtag that appears in the tweet. Other hashtags are ignored.

True
False
False → #lol
True
⋮

- List one advantage and one disadvantage of using k-Nearest Neighbours for this specific problem and explain your answer.
- ~~List one advantage and one disadvantage of using Naive Bayes for this specific problem and explain your answer.~~

[6 marks]

+ve

- Good at approximating complex functions
- Can be used for non-linear data
- Simple distance metric can be used (T/F)

-ve

- Singular words may change the meaning a lot e.g. "not", "never"
- High computational cost (computing for every point)
- Space/memory-intensive

Shay

2) a) One advantage of using kNN for this problem is that we can easily see which tweets will be similar based on whether or not a tweet contains similar words to another tweet.

One disadvantage of using kNN is that it is likely to be computationally expensive, computing distances to every tweet in the dataset every time we want to classify a new tweet.

Adwit

2a) • We defer all computation until prediction as we don't have to train a model
• KNN doesn't fair well with high dimensional data due to its complexity being O(nd), needs to calculate the distance from every point to a new one for every feature.

Question 3 Linear Regression and Logistic Regression

(a) The cost of a hypothesis function parametrised by z is given by the following equation:

$$z^2 - 12z + 2$$

What is the value of the parameter z at which the cost is minimum?

[4 marks]

Shay

3) a) $z^2 - 12z + 2 = f(z)$

$f'(z) = 2z - 12$ $\qquad f''(z) = 2$

At a minimum, $f'(z) = 0$ and $f''(z) > 0$

$2z = 12$ $\qquad 2 > 0$ ✓

$\underline{\underline{z = 6}}$

Adwit

3a) $\cancel{z}cost(z) = z^2 - 12z + 2$

$(z-6)^2 + 2 - 36 = (z-6)^2 - 34$

Minimum cost z-value: $\underline{\underline{z = 6}}$

(b) The cost function for Logistic Regression is given by the following equation:

$$Cost(w) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} log(h_w(x^{(i)})) + (1 - y^{(i)}) log(1 - h_w(x^{(i)}))\right]$$

where w represents the weights of the hypothesis function h, and $y^{(i)}$ and $x^{(i)}$ are the input and output values of a given example.

Derive this expression from the version which shows the cost for each case $y = 0$ and $y = 1$ separately. Additionally, detail why this is a reasonable cost function. You might find it easier to use the separated version to show this by analysing the values of $-log(x)$ and $-log(1 - x)$.

[8 marks]

Adwit

b) Separated form: $cost(z) = \begin{cases} -Log(h_w(x^{(i)})) & \text{if } y=1 \\ -log(1-h_w(x^{(i)})) & \text{if } y=0 \end{cases}$

• When we write it as 1 function, we ensure the multiplier of the one we need to use is 1 and the multiplier is 0 for the other term in summation.

$-\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} Log(h_w(x^{(i)})) + (1-y^{(i)}) Log(1-h_w(x^{(i)}))\right]$

When $y=1$ the multiplier of $Log(h_w(x^{(i)}))$ is 1 whereas multiplier for $Log(1-h_w(x^{(i)}))$ is 0 as $1-1=0$

$$L_{CE} = \begin{cases} -\log(z) & \text{if } y=1 \\ -\log(1-z) & \text{if } y=0 \end{cases}$$

where z is the model output

(c) The Hypothesis function for Univariate Linear Regression is $y = w_0 + w_1 x$. The cost function associated with this hypothesis function h, parametrised by some w_0 and w_1, is $\sum_i (y^{(i)} - h_w(x^{(i)}))^2$, where $y^{(i)}$ and $x^{(i)}$ represent the output and input values of the i$^{\text{th}}$ training example.

(i) Write out and provide an explanation for the general form of the hypothesis function and cost function of Linear Regression in two variables.

(ii) Similarly, write out and provide an explanation for the general form of the hypothesis function for Univariate Non-linear Regression. Assume that the non-linear hypothesis function is a quadratic.

[8 marks]

Shay

c) i) The hypothesis function $h(x) = w_0 + w_1 x$ is defined in this way as it creates a line in 2-dimensional space relating x values to predictions. w_0 represents the y-offset or intercept of the line, and w_1 represents the gradient of the line.

The cost function (mean squared error / L2 loss) simply computes the difference between each actual data label and the model's prediction, and squares it (eliminating negative values by making them positive). This is done for every training example and then added up to get our cost function value.

ii) $k: y = w_0 + w_1 x^1 + w_2 x^2$

We have a hypothesis function k which just takes in 1 x value, and then we use that raised to different powers to get a polynomial function.