Georgia Institute of Technology

CS 6601 Artificial Intelligence

Final Exam

Fall 2024 - **Solutions**

Duration of Exam: December 2nd 2024, 8:00 AM (EDT) - December 9th 2024, 7:59 AM (EDT)

Weight: 20% **Points:** 100 + 3 Extra Credit

Submission: Via Canvas. Unlimited attempts during exam window, **only last submission will be used for grading**.

Clarifications and Corrections

Clarifications have been added in red below. Corrections have been added in purple below and in the questions themselves.

Search

12/2/24 - Treat $h(n)$ as the heuristic from state n to the goal. We've already provided the $h(n)$ values for you so you should not need to do any calculations to get the heuristic.

Regardless of what the goal is, $h(n)$ returns:

- h(earth stations) = 30
- h(mars stations) = 20
- h(jupiter stations) = 10
- h(saturn stations) = 0

12/2/24 - Do not assume that the A* priority queue breaks ties by FIFO like in A1.

12/3/24 - Search Q3 PDF is missing some answer options. Use the Canvas answer options.

12/3/24 - The goal state is S3 in Q1, Q2, and Q4. In Q3 the goal is variable depending on which answer option you are investigating. $h(n)$ is a heuristic - it is a function that returns some value that you use in $f(n)$, and we have provided that value for you for each station. e.g. the value of the heuristic from E1 to any goal when using $h(E1)$ is 30.

12/7/24 - Q1 should be **Select One**, not Multiple Select

Not announced but will be addressed in regrading - Q4 is worth 2 points as shown in the PDF. It was erroneously marked as 1 point on the exam.

Optimization

12/2/24 - Figure 6 is incorrect in both PDF and Canvas exam. Correct figure is shown below. Also note that swap for mutation doesn't have to be adjacent like the example, can be any two cells.

6	5	4
7	8	1
9	3	2

(a) Before mutation

6	4	5
7	8	1
9	3	2

(b) After mutation

Figure 1: Mutation where the digits 5 and 4 are swapped.

12/2/24 - The square presented in Figure 4 has fitness score of 31, not 26. Table 3 is still correct, the result of the summation is wrong.

12/2/24 - Table 2's rows/columns/diagonals are mixed up. It should look as follows:

Row Sums	**Column Sums**	**Diagonal Sums**
$8+1+6=15$	$8+3+4=15$	$8+5+2=15$
$3+5+7=15$	$1+5+9=15$	$6+5+4=15$
$4+9+2=15$	$6+7+2=15$	

Table 1: Row, Column, and Diagonal Sum calculations for Figure 3

Game Playing

12/3/24 - Quadrants: There are 4 Quadrants. Top Left 2x2, Top Right 2x2, Bottom Left 2x2, and Bottom Right 2x2.

12/6/24 - 3.2 Game State: The question is asking for the number unique board states considering Player 1's first move. The boards shown in Figure 14 are examples of such board states.

12/6/24 - Defining "Depends exponentially on": If we have a function f, such that $f = a^b$, we would say that f depends exponentially on a and that f depends exponentially on b.

CSPs

Not released publicly but most frequently asked - Why is the number of complete assignments 8^8 and not 3^8 when we make the variables the ports? 3^8 comes from the domains of each port being $\{S, M, L\}$. But this does not adequately represent the states that each port can be assigned as it doesn't account for the five other possible states of $\{\text{No Ships}, S \text{ and } M, S \text{ and } L, M \text{ and } L, \text{All 3 ships}\}$

Probability

12/2/24 - Probability Q4 (Major Change) - Please ignore the hint "Hint: The answer is a whole number". The problem should instead say: "The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p+q$. Hint: The absolute value of $p-q$ is 25." Answer as a whole number

12/4/24 - Q3: "The highest number that is encountered during the spin" includes the last spin. In other words, you will always get at least $4 when the game ends. Here's 3 scenarios:

1. 1, 1, 3, 2, 4: You get $4.
2. 6, 2, 1, 5, 5, 4: You get $6.
3. 1, 5, 3, 4: You get $5.

12/6/24 - i.i.d = Independent and identically distributed

Bayes Nets

12/2/24 - Q4 typo: (D = Effective) should instead say (R = Effective)

12/2/24 - Q5 typo: (A = Increased) should instead say (A = Triggered)

Machine Learning

12/3/24 - Please use bits for entropy calculation unless specified otherwise.

12/7/24 - We are building a decision tree to classify fish species.

12/7/24 - Machine Learning Extra Credit, additional reading: LINK

Pattern Recognition Through Time

12/2/24 - Pattern Recognition Through Time Q8 (Major Change) - Please ignore the statement "You can either start from the top right or the bottom left cell". The statement should instead say:

"You should start from the top right cell and traverse to the bottom left cell." DO NOT start from the bottom left. The rest of the instructions remain the same and are applicable as long as you start from the top right cell and traverse to the bottom left cell while finding the path.

12/5/24 - For the HMM Trellis, the expected values are finite decimals, and can be calculated by hand. Be aware of any floating point precision errors. You can try calculating the pre-filled values using your approach to cross-check if it is correct.

12/5/24 - If you are confused about how to find the path for Dynamic Time Warping, the path is identified the same way as mentioned in Challenge Question 8: "This path is identified by starting with the the cell at the top right and moving towards the bottom left diagonally, left, or down at each step, choosing the smallest value among the options. In the event of a tie between the values, move diagonally." However, unlike CQ8, the distance between the two songs is given by the value in the top right cell.

Logic and Planning

12/2/24 - In part 2 of the question, predicate "not Cargo(c)" means "Cargo c doesn't need to be picked."

12/4/24 - In Part 2, consider "valid actions" to be any actions that does not violate the constraints.

12/4/24 - In part 1, assume 0 = False and 1 = True in the truth table.

12/4/24 - In part 2, assume the facilities and the cargo center can hold infinite amount of cargo.

12/5/24 - For Q 9.3 & 4, consider the question is asking for the most complete formalization of the states.

Planning Under Uncertainty

12/2/24 - Packages A, B, and C correspond to the small, medium, and large package respectively.

12/2/24 - Q7 - Do not consider package C (assume package C not an option)

12/4/24 - Q5 - Assume the statements can occur in isolation

12/6/24 - Q10 - Make selections that are definable and have some semblance of consistency

Contents

1 Search [5pts]

Author: Prajwal Mohan Kumar

Introduction

In the year 2150, space stations scattered across the Solar System rely on a central supply station located on Earth for resources. Your mission is to develop a navigation system for an autonomous spacecraft that needs to deliver critical supplies from Earth to a given space station while minimizing travel cost. The spacecraft can travel between planets and space stations using pre-determined space routes, and the travel cost (fuel) is associated with the distance between stations. The space routes between the stations are illustrated in the graph shown in Figure 2 below.

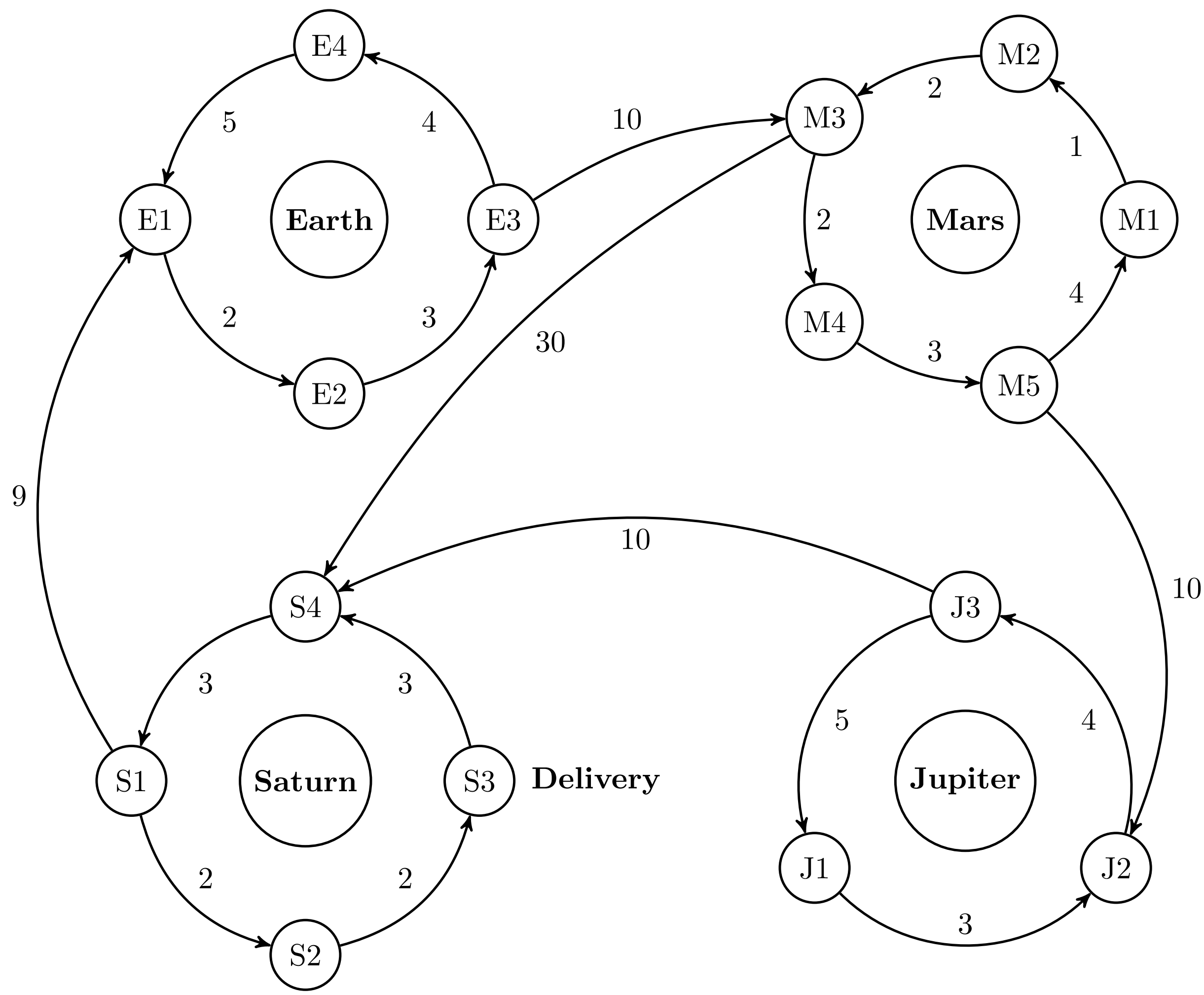


Figure 2: Space Routes

Note that each planet is represented by a central node with its space stations orbiting around it. Arrows indicate space routes between stations and are directed edges in the graph (e.g. you can go from M5 to J2 directly but not vice versa). All routes have a fuel cost as indicated by the graph unless specified otherwise.

The spacecraft starts at Station E1 orbiting Earth and needs to deliver supplies to the Station S3 orbiting Saturn. The goal is to find the optimal path for the spacecraft to minimize fuel consumption while ensuring the delivery is completed. Stations may be visited more than once.

Heuristic Function: Let us define our heuristic function $h(n)$ for A* as a pre-calculated estimated fuel expenditure from a given station's planetary system to the planet Saturn. The heuristic values are assigned as follows:

- Stations in the **Earth** system (E1–E4): $h(n) = 30$
- Stations in the **Mars** system (M1–M5): $h(n) = 20$
- Stations in the **Jupiter** system (J1–J3): $h(n) = 10$
- Stations in the **Saturn** system (S1–S4): $h(n) = 0$

More specifically each station's heuristic value can be seen in Table 2 below.

Station	E1	E2	E3	E4	M1	M2	M3	M4	M5	J1	J2	J3	S1	S2	S3	S4
Heuristic $h(n)$	30	30	30	30	20	20	20	20	20	10	10	10	0	0	0	0

Table 2: $h(n)$ values for each station

1.1 Search Q1 [1pt]

Is the heuristic $h(n)$ admissible and/or consistent in this problem? Select all that apply.

Select one.

☒ The heuristic $h(n)$ is admissible and consistent.

☐ The heuristic $h(n)$ is admissible but NOT consistent.

☐ The heuristic $h(n)$ is consistent but NOT admissible.

☐ The heuristic $h(n)$ is neither admissible nor consistent.

Solution: Note that the search is from E1 to S3.

Admissibility: $h(n) \leq g(n)$

- E1: $h(n) = 30 \leq g(n) = 51$
- E2: $h(n) = 30 \leq g(n) = 49$
- E3: $h(n) = 30 \leq g(n) = 46$
- E4: $h(n) = 30 \leq g(n) = 56$
- M1: $h(n) = 20 \leq g(n) = 39$
- M2: $h(n) = 20 \leq g(n) = 38$
- M3: $h(n) = 20 \leq g(n) = 39$
- M4: $h(n) = 20 \leq g(n) = 37$
- M5: $h(n) = 20 \leq g(n) = 34$
- J1: $h(n) = 10 \leq g(n) = 24$
- J2: $h(n) = 10 \leq g(n) = 21$
- J3: $h(n) = 10 \leq g(n) = 17$

- S1: $h(n) = 0 \leq g(n) = 4$
- S2: $h(n) = 0 \leq g(n) = 2$
- S3: $h(n) = 0 \leq g(n) = 0$
- S4: $h(n) = 0 \leq g(n) = 7$

Consistency: $h(n) \leq g(n, n') + h(n')$. For consistency to fail, we need to find a case where $h(n) > g(n, n') + h(n')$. We also can recognize that the $h(n)$ value for stations are the same for each planet so we can just investigate planet to planet cases instead of investigating all possible pairs of stations.

$h(\text{Earth}) \overset{?}{\leq} g(\text{Earth}, \text{Mars}) + h(\text{Mars})$ - To make this statement false, we need as small of a $g(\text{Earth}, \text{Mars})$ value as possible. The min(g(Earth, Mars)) = 10 and this value still satisfies the consistency requirement $(30 \leq 10 + 20)$. We can follow this approach for the pairs of planets.

$h(\text{Earth} \overset{?}{\leq} g(\text{Earth}, \text{Jupiter}) + h(\text{Jupiter}) \rightarrow \min(g(\text{Earth}, \text{Jupiter})) = 25 \rightarrow 30 \leq 25 + 10$

$h(\text{Earth} \overset{?}{\leq} g(\text{Earth}, \text{Saturn}) + h(\text{Saturn}) \rightarrow \min(g(\text{Earth}, \text{Saturn})) = 35 \rightarrow 30 \leq 35 + 0$

$h(\text{Mars} \overset{?}{\leq} g(\text{Mars}, \text{Jupiter}) + h(\text{Jupiter}) \rightarrow \min(g(\text{Mars}, \text{Jupiter})) = 10 \rightarrow 20 \leq 10 + 10$

$h(\text{Mars} \overset{?}{\leq} g(\text{Mars}, \text{Saturn}) + h(\text{Saturn}) \rightarrow \min(g(\text{Mars}, \text{Saturn})) = 24 \rightarrow 20 \leq 24 + 0$

$h(\text{Jupiter} \overset{?}{\leq} g(\text{Jupiter}, \text{Saturn}) + h(\text{Saturn}) \rightarrow \min(g(\text{Jupiter}, \text{Saturn})) = 10 \rightarrow 10 \leq 10 + 0$

Since no contradictions were found, the heuristic is also consistent.

1.2 Search Q2 [1pt]

Let us use the A* algorithm using the heuristic $h(n)$ to calculate the optimal delivery route from E1 to S3. Provide the path found by the A* algorithm.

Answer as a sequence of stations (fill in the blank).

Answer: $\boxed{\text{E1} \rightarrow \text{E2} \rightarrow \text{E3} \rightarrow \text{M3} \rightarrow \text{M4} \rightarrow \text{M5} \rightarrow \text{J2} \rightarrow \text{J3} \rightarrow \text{S4} \rightarrow \text{S1} \rightarrow \text{S2} \rightarrow \text{S3}}$

Solution: This is the optimal path found by applying A* while using the given $h(n)$. It helps to note that $h(n)$ being admissible and consistent and $h(\text{goal}) = 0$ guarantees that A* will find the optimal path.

1.3 Search Q3 [1pt]

Command wants us to calculate the optimal path to any space station. Unfortunately there is no budget to pre-calculate new heuristics between planetary systems so we have to make use of our existing heuristic $h(n)$. For which of the following goals will usage of the $h(n)$ heuristic for A* guarantee the return of the optimal path to that goal? Note that the starting station is still E1. Select all that apply.

Multiple select.

☒ E2

☒ E3

☒ E4

☒ M1

☒ M2

☒ M3

☒ M4

☒ M5

☒ J1

☒ J2

☒ J3

☒ S1

☒ S2

☒ S4

Solution: Note that while the heuristic $h(n)$ may no longer be admissible for all of these goals, the structure of the graph makes it such that the optimal path from $E1$ to all of these goals is still the path find by A*.

1.4 Search Q4 [2pts]

Suppose a hacker secretly changes the heuristic function $h(n)$ for Stations J1, J2, and J3 by increasing their heuristic values by +10. How does this adjustment impact the A* search and the path found by A*? Select all that apply.

Multiple select.

☐ The $f(n)$ values at J1, J2, and J3 will remain the same.

☐ The overall path found by A* will remain the same.

☒ The cost of the overall path found by A* will increase.

☒ The heuristic adjustment will change the order in which nodes are explored in the A* search.

Solution:

- The $f(n)$ values at J1, J2, and J3 will increase. The $g(n)$ values to reach J1, J2, and J3 will more or less stay the same while the $h(n)$ values will increase by 10. Since $f(n) = g(n) + h(n)$, $f(n)$ will change in value.
- The overall path found by A* will not remain the same. The $f(n)$ cost of the path E1 → E2 → E3 → M3 → S4 is 45 + 0. The $f(n)$ cost of the path E1 → E2 → E3 → M3 → M4 → M5 → J2 is $35 + 20 = 55$ which is greater than the $f(n)$ to S4, which means that the path to S4 will be explored by the algorithm before the path to J2. From there, the path through S4 to S3 will be $f(n) = 52$ which is still lower than 55, so a non-optimal path to S3 will be found before the optimal path can be explored.
- The cost of the overall path found by A* will increase due to a non-optimal path being returned by A* per the conclusion above.
- Similarly the heuristic adjustment has changed the order in which nodes are explored in the A* search.

2 Optimization [6pts]

Author: Rory McGurty

Introduction

A magic square is defined as a $n \times n$ square array of the numbers 1, 2 ... n^2 where each row, column and diagonal sum to the same value. A magic square of size $n = 3$ is shown in Figure 3, and its rows, columns, and diagonals sum to 15, the target sum.

8	1	6
3	5	7
4	9	2

Figure 3: Sample valid 3×3 magic square with row, column, and diagonal sum of 15

Row Sums	Column Sums	Diagonal Sums
$8 + 1 + 6 = 15$	$8 + 3 + 4 = 15$	$8 + 5 + 2 = 15$
$3 + 5 + 7 = 15$	$1 + 5 + 9 = 15$	$6 + 5 + 4 = 15$
$4 + 9 + 2 = 15$	$6 + 7 + 2 = 15$	

Table 3: Row, Column, and Diagonal Sum calculations for Figure 3

For a size 3 magic square, there are a total of 8 solutions. Note that we consider rotations and reflections as distinct magic squares. While finding a solution for a size 3 magic square is quite trivial, the problem of finding a solution for magic squares of larger sizes is a more complex problem in combinatorics that often requires the use of optimization algorithms. We will stick with a magic square of size 3 in this problem but develop an idea of how we can use evolutionary algorithms to find a solution.

Let us define a **valid** square as a square in which no digit appears more than once. An example comparing a valid and an invalid square can be seen in Figure 4. In this optimization problem, the evolutionary algorithm is only allowed to generate valid squares, any invalid squares generated will be removed and not considered in the problem.

6	5	4
7	8	1
9	3	2

(a) A **valid** square

2	2	8
6	1	9
5	3	7

(b) An **invalid** square

Figure 4: Note that (b) is invalid because the digit 2 appears twice

2.1 Optimization Q1 [1pt]

Consider an algorithm that randomly generates **valid** squares with replacement (it may generate the same square more than once). What is the expected value for the number of attempts it will take for this algorithm to find a solution? If necessary, round your answer to the nearest whole number.

Answer as a whole number.

Answer: $\boxed{45360}$

Solution: The probability p that the algorithm randomly generates a solution at any given round is:

$$p = \frac{\text{Number of solutions}}{\text{Number of valid squares}}$$

$$p = \frac{8}{9!} = \frac{1}{45360}$$

The expected value is then $1/p = 45360$.

2.2 Optimization Q2 [1pt]

Let us define a fitness function for calculating the fitness score of any given valid square. Our fitness function is defined as the sum of the absolute difference between each row sum, column sum, diagonal sum, and the target sum – 15 for a magic square of size $n = 3$. We use our fitness score to evaluate how close a given valid square is to a magic square solution, with a magic square solution having a fitness score of zero (a smaller fitness score value is better). A sample valid square is shown in Figure 5 and it has a fitness score of 31 as calculated using Table 4 below.

1	3	8
2	4	7
9	6	5

Figure 5: Sample valid square with fitness score of 31

Row Diffs	Column Diffs	Diagonal Diffs
$\mid 1+3+8-15 \mid = 3$	$\mid 1+2+9-15 \mid = 3$	$\mid 1+4+5-15 \mid = 5$
$\mid 2+4+7-15 \mid = 2$	$\mid 3+4+6-15 \mid = 2$	$\mid 8+4+9-15 \mid = 6$
$\mid 9+6+5-15 \mid = 5$	$\mid 8+7+5-15 \mid = 5$	

Table 4: Fitness score $= 3+2+5+3+2+5+5+6 = 31$

What is the fitness score for the square shown in Figure 6 below?

7	9	5
3	6	2
8	1	4

Figure 6: Calculate the fitness score for this square

Answer as a whole number.

Answer: $\boxed{26}$

Solution:

Row Diffs	**Column Diffs**	**Diagonal Diffs**
$\mid 7+9+5-15 \mid = 6$	$\mid 7+3+8-15 \mid = 3$	$\mid 7+6+4-15 \mid = 2$
$\mid 3+6+2-15 \mid = 4$	$\mid 9+6+1-15 \mid = 1$	$\mid 5+6+8-15 \mid = 4$
$\mid 8+1+4-15 \mid = 2$	$\mid 5+2+4-15 \mid = 4$	

Table 5: Fitness score $= 6+4+2+3+1+4+2+4 = 26$

2.3 Optimization Q3 [1pt]

An important step in Genetic Algorithms involves mutations. We define mutations in this problem as a random swapping of two digits in a valid square. No more than one swap can occur per mutation. An example of a mutation can be seen in Figure 7 below.

6	5	4
7	8	1
9	3	2

(a) Before mutation

6	4	5
7	8	1
9	3	2

(b) After mutation

Figure 7: Mutation where the digits 5 and 4 are swapped.

Given the valid square in Figure 8, what is the **best** fitness (smallest fitness score value) that can be obtained from the square with one mutation?

1	2	3
4	5	6
7	8	9

Figure 8: Square for Q3 and Q4

Answer as a whole number.

Answer: $\boxed{12}$

Solution:

Swap	Fitness Score	Swap	Fitness Score
No Swap	24		
1 – 2	25	4 – 5	26
1 – 3	24	4 – 6	20
1 – 4	27	4 – 7	27
1 – 5	30	4 – 8	26
1 – 6	27	4 – 9	27
1 – 7	24	5 – 6	26
1 – 8	25	5 – 7	30
1 – 9	12	5 – 8	30
2 – 3	25	5 – 9	30
2 – 4	28	6 – 7	27
2 – 5	30	6 – 8	28
2 – 6	26	6 – 9	27
2 – 7	29	7 – 8	25
2 – 8	12	7 – 9	24
2 – 9	25	8 – 9	25
3 – 4	27		
3 – 5	30		
3 – 6	27		
3 – 7	24		
3 – 8	29		
3 – 9	24		

Table 6: Q3 Swap Scores

2.4 Optimization Q4 [1pt]

Given the valid square in Figure 8, what is the **worst** fitness (largest fitness score value) that can be obtained from the square with one mutation?

Answer as a whole number.

Answer: $\boxed{30}$

Solution: See Table 6

2.5 Optimization Q5 [1pt]

Our Genetic Algorithm follows the following procedure:

1. Generate a valid square at random
   (a) If the valid square is a magic square solution, stop and return the square.
   (b) Otherwise, the valid square is NOT a magic square solution – generate a new valid square by randomly swapping two elements of the current square (a mutation).
      i. If the fitness score of the new square is better than the current square, set the current square = new square.
      ii. Otherwise the current square does not change.
2. Perform as many iterations as needed until a magic square solution is found.

This algorithm is not guaranteed to find a solution. Which of the following changes can be applied such that applying the change alone can guarantee that a solution will be found by the algorithm? If answer choice X requires answer choice Y to also be selected in order to guarantee a solution, do not select answer choice X. Select all that apply.

Multiple select.

☒ Use random restart. If the algorithm cannot find a solution with one starting square permutation, try another starting square permutation.

☒ Allow a 10 percent chance that a square permutation can transition to a less fit state.

☐ In addition to a random swap, try a random rotation as well.

☐ You can now either choose to swap two grid positions, two grid rows, or two grid columns. Only one of these swaps can be performed each iteration.

Solution:

1. Option 1. Yes. Random restart will eventually find a solution. Given enough time the solution grid will randomly be chosen as the starting grid.Option 2. Yes. The randomness allows the algorithm to escape any local minima to find the global minimum.
2. Option 2. Yes. The randomness allows the algorithm to escape any local minima to find the global minimum.
3. Option 3. No. A random rotation returns the same grid with the same fitness score.
4. Option 4. No. This example grid has a score of 2 and neither a swap of grid places, rows, or columns will decrease the fitness score.

8	1	7
4	5	6
3	9	2

Figure 9: Option 4 Example

2.6 Optimization Q6 [1pt]

Let us consider different crossover schemes for this problem.

First, we define a **valid** crossover scheme to be one where the crossover produces children that are **valid** squares (no digits appears more than once).

A	**B**	**C**	**D**	**E**	**F**

(a) Parent 1

G	H	I	J	K	L

(b) Parent 2

Figure 10: Parents

Next we introduce two crossover schemes, the **two-point crossover** scheme and the **uniform crossover** scheme. The examples below use the parents in Figure 10 above.

In the **two-point crossover** scheme each parents has two points where crossover occurs. In the example shown in Figure 11 below, AB and EF from the first parent are combined with IJ from the second parent, while GH and KL from the second parent are combined with CD from the first parent to produce the two new children.

A	**B**	I	J	**E**	**F**

(a) Child 1

G	H	**C**	**D**	K	L

(b) Child 2

Figure 11: Two-Point Crossover

In the **uniform crossover** scheme the crossover alternates between parents for every cell. An example of a uniform crossover between Parent 1 and Parent 2 is shown below.

A	H	**C**	J	**E**	L

(a) Child 1

G	**B**	I	**D**	K	**F**

(b) Child 2

Figure 12: Uniform Crossover

Assume we want to improve our algorithm and include crossover in addition to mutation. Select all of the **valid** crossover schemes below. Assume a constant population size of 10 where parents are chosen proportional to their fitness and every child has only two parents. Select all that apply.

Multiple select.

☐ Two-point crossover. The first and third row of one parent is combined with the second row of the other parent.

☐ Uniform crossover. A child is created by alternately choosing from each parent. The new square will be composed of 5 values from one parent and 4 from the other.

☐ The children are produced by a uniform crossover that works from left to right and top to bottom. If a value already exists in the child (e.g. a 4 is chosen from one parent to pass onto the child but it is already present in the child from another parent) the value at that same position from the other parent is used instead.

☒ A child has the first two rows of one parent. Then, the third row of the other parent is added one by one to the third row of the child. If any of the values in the third row also appear somewhere in the first two rows of the child, a number that hasn't appeared in the child yet is used instead.

Solution:

Option 1. Two parents create an invalid child with duplicate rows.

2	9	1
7	8	3
6	5	4

(a) Parent 1

2	9	1
6	5	4
7	8	3

(b) Parent 2

2	9	1
6	5	4
6	5	4

(c) Invalid Child

Figure 13: Option 1

Option 2. Two parents create an invalid child with two 8s.

2	9	1
7	8	3
6	5	4

(a) Parent 1

2	9	1
8	7	3
6	5	4

(b) Parent 2

2	9	1
8	8	3
6	5	4

(c) Invalid Child

Figure 14: Option 2

Option 3. Two parents create an invalid child with two 9s. The scheme does not select the 2 from the second parent to pass on to the child because it has already passed on a 2 from the first parent. It replaces the 2 from the second parent with a 9 from the first parent. However, another 9 is passed on to the child from the second parent as well.

2	9	1
7	8	3
6	5	4

(a) Parent 1

1	2	7
9	8	3
6	5	4

(b) Parent 2

2	9	1
9	8	3
6	5	4

(c) Invalid Child

Figure 15: Option 3

Option 4. This option will produce valid children.

3 Game Playing [7pts]

Author: Adam Mazlout

Introduction

Two players are playing a board game called Quantik, where White always plays first (Player 1). There are four types of pieces: Circle, Square, Pentagon, and Triangle, as seen in Figure 16.

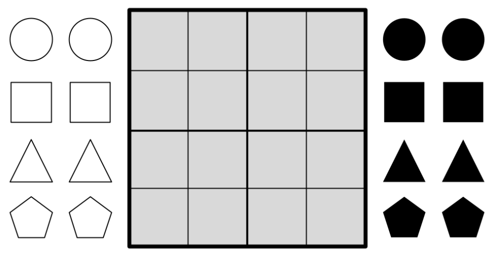
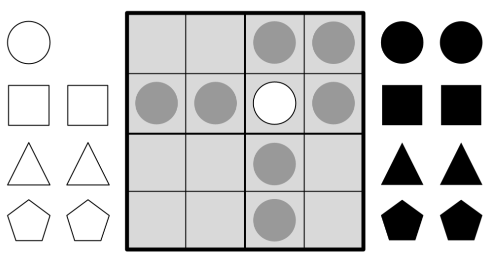
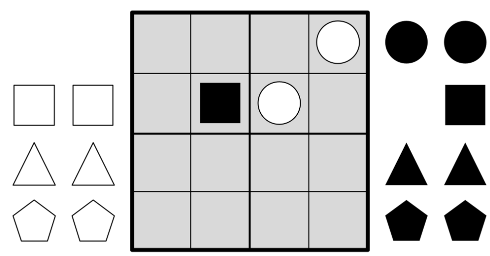


Figure 16: Quantik board and pieces

Each player has two of each type of piece, for a total of 8 pieces per player. Players take turns placing a piece on the board. At a given player's turn, they can only place a piece in a spot where no opponent's piece of the same type is in the same row, column, or quadrant, as shown in Figure 17.



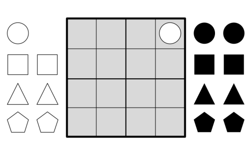
(a) Locations where black cannot place a circle after white placed their circle



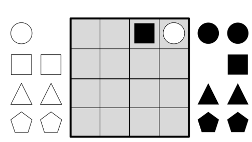
(b) Two white circles in the same quadrant. Players pieces do not block their own pieces from being placed

Figure 17: Quantik rules

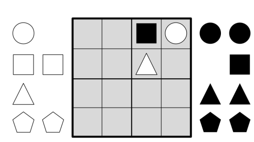
A player wins if they are the first to place the 4th piece to completely fill a row, column, or quadrant. See Figure 20 below for an example.



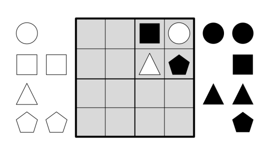
(a) Turn 1: White places circle



(b) Turn 2: Black places square



(c) Turn 3: White places triangle



(d) Turn 4: Black places pentagon and wins the game

Figure 18: Sample Quantik game where Black wins by filling a quadrant

Let us define the evaluation function for determining the utility value of a game state as the maximum possible moves available to a player in a given game state. For example, from the perspective of Player 1, an empty board has a utility value of $4 \cdot 16 = 64$ since each type of piece has 16 possible positions for placement.

3.1 Game Playing Q1 [1pt]

What is the utility value of the game state for Player 2 after Player 1 has placed down one piece?

Answer as a whole number.

Answer: $\boxed{53}$

Solution: The board will look something like Figure 17 (a). The utility value is calculated as $3 \cdot 15 + 8 = 53$.

3.2 Game Playing Q2 [1pt]

Notice that from an empty board, there are 64 possible game states that can result from Player 1's first move. In order to reduce the search space and speed up our game playing algorithm, we can take advantage of symmetry properties of Quantik game states to reduce equivalent game states by symmetry. If we take into account **rotational symmetry about the center of the board, reflective symmetry across the vertical/horizontal/diagonal axes, and substitution symmetry (where piece types can be substituted for each other as shown in Figure 19)**, how many unique game states are there after Player 1 makes the first move on an empty board?

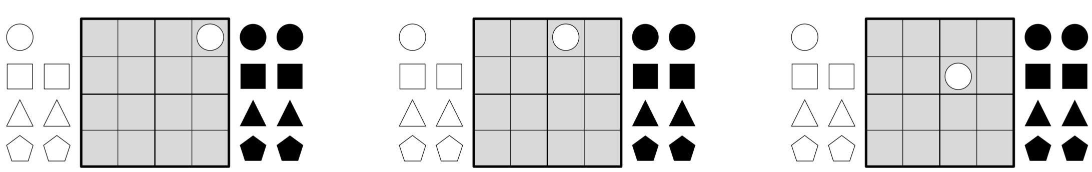
(a) State X (b) State Y

Figure 19: State X and State Y are equivalent by substitution symmetry

Answer as a whole number.

Answer: $\boxed{3}$

Solution:



(a) (b) (c)

Figure 20: Three unique Quantik board states

Player 1 is trying to decide what their best first move is for Quantik. They decide to make a Game Tree of depth 3 to help evaluate their options (for simplicity they have ignored some nodes). The game tree is shown in Figure 21 below, use this game tree for Q3-Q6.

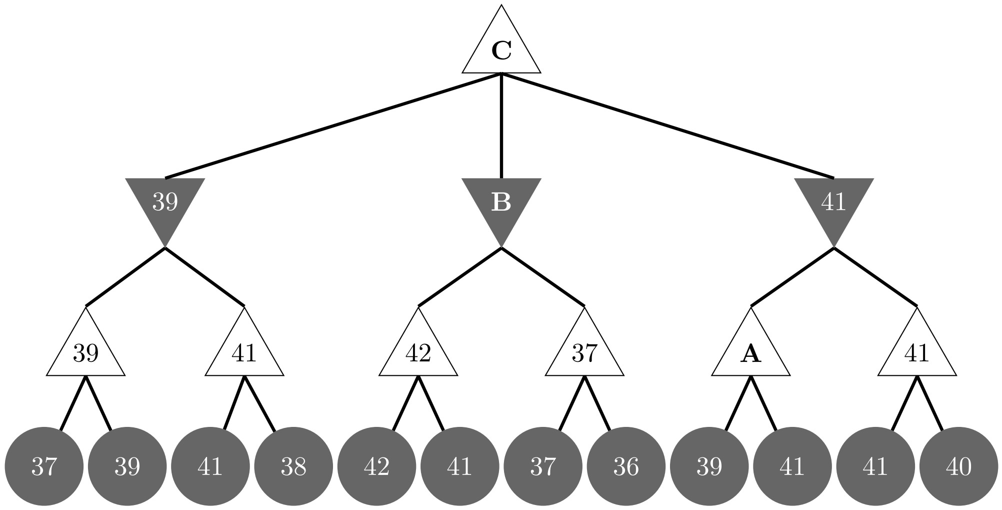


Figure 21: Game Tree

3.3 Game Playing Q3 [1pt]

Using minimax to propagate utility values up the tree, what is the utility value of **node A**?

Answer as a whole number.

Answer: $\boxed{41}$

Solution: $\max(39, 41) = 41$

3.4 Game Playing Q4 [1pt]

Using minimax to propagate utility values up the tree, what is the utility value of **node B**?

Answer as a whole number.

Answer: $\boxed{37}$

Solution: $\min(42, 37) = 37$

3.5 Game Playing Q5 [1pt]

Using minimax to propagate utility values up the tree, what is the utility value of **node C**?

Answer as a whole number.

Answer: $\boxed{41}$

Solution: $\max(39, B, 41) = \max(39, 37, 41) = 41$

3.6 Game Playing Q6 [1pt]

How many leaf nodes will be pruned if we apply Alpha-Beta pruning from left to right to the Game Tree in Figure 21? Assume that the maximum possible utility is 42 and the minimum possible utility is 36.

Answer as a whole number.

Answer: $\boxed{3}$

Solution:

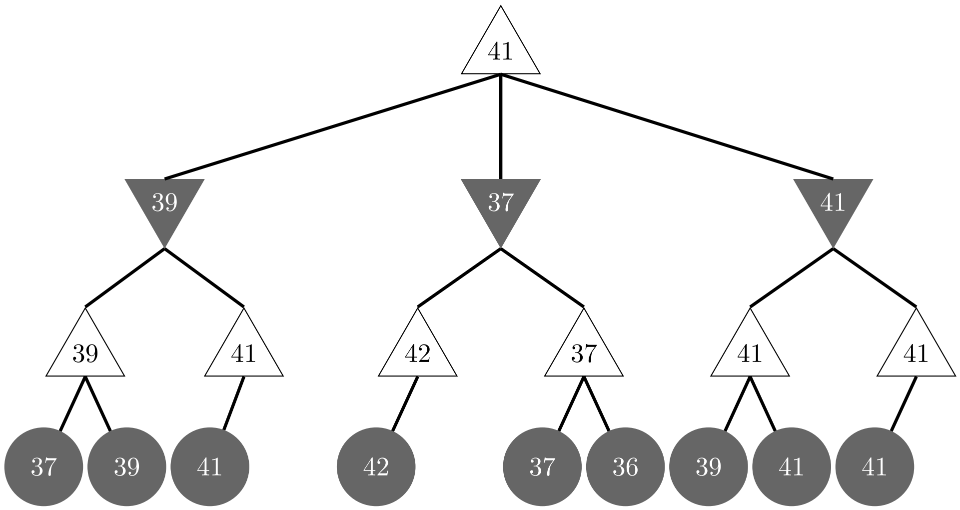


Figure 22: Alpha-Beta Pruned Game Tree

3.7 Game Playing Q7 [1pt]

In general, how does minimax compare to alpha-beta pruning in space and time complexity? Select all that apply.

Multiple select.

- ☒ Alpha-beta pruning's time complexity is always better than or equivalent to minimax's
- ☐ Alpha-beta pruning always has a higher space-complexity since it needs to keep track of the alphas and betas.
- ☐ Minimax's time complexity depends linearly on the depth
- ☒ Minimax's time complexity depends exponentially on the branching factor
- ☒ For smaller, shallower game trees, minimax is faster than alpha-beta due to lower overhead.

Solution:

1. A is true because alpha-beta pruning reduces the number of nodes explored by skipping branches, thus, reducing time-complexity. In the worst case, it is equivalent to minimax if the nodes are ordered such that no nodes can be pruned.
2. B is false because alpha-beta pruning does not impact space complexity so they both have the same space complexity of $O(bd)$.

3. C is false because minimax's time complexity is $O(b^d)$. Ad d increases, the size of the tree grows exponentially, not linearly.
4. D is true because minimax's time complexity is $O(b^d)$. As b increases, the size of the tree grows exponentially.
5. E is true because alpha-beta pruning adds some computational overhead, and for smaller game trees, the overhead may outweigh the benefits of pruning.

4 Constraint Satisfaction Problems (CSPs) [6pts]

Author: Raymond Jia

Introduction

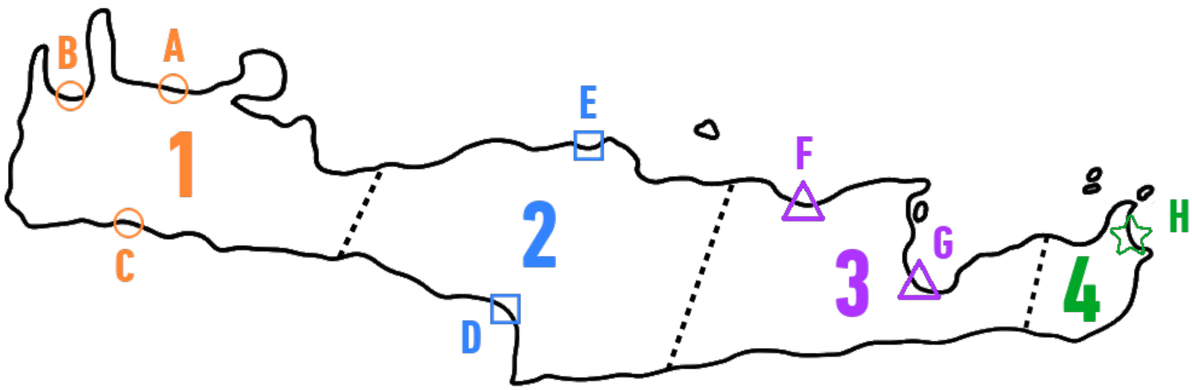


Figure 23: The Ports of Christmas Island

As the shipping coordinator for Christmas Island, you've been informed that three ships will be arriving simultaneously on Christmas Eve to deliver toys. Checking your schedule, you see that there will be no other ships at the island on Christmas Eve, so you just have to assign ports for the three ships to dock at. Unfortunately it isn't as easy as just assigning any ports to the three ships, it turns out there are some constraints that we have to account for!

Variables: There are 3 **variables**, and they are the 3 **ships**. We will refer to them using letters as follows: **S** - the small ship, **M** - the medium ship, and **L** - the large ship.

Domains: For each variable, there are **8 values** that can be assigned and those values are the eight ports that they can dock at, which we will refer to using the letters **A through H**.

With our variables and their domains defined, we can now take a look at the constraint requirements. To save you time, we've collected the unary and binary constraints in table format. If you want to see the requirements that brought us to these constraints, please take a look at Appendix A [12].

Unary Constraints

1. The large ship (**L**) cannot dock at port **D** or **F**. (Because it is too big!)

Continue to next page →

Binary Constraints

We've reduced the binary constraints into allowed pairs format between each variable for your ease of use.

Edge	Total	Allowed Tuples
(**L**arge, **M**edium)	35	(A, A), (A, B), (A, C), (A, D), (A, E)
		(B, A), (B, C), (B, D), (B, E)
		(C, A), (C, B), (C, D), (C, E)
		(D, A), (D, B), (D, C), (D, E), (D, F), (D, G)
		(E, A), (E, B), (E, C), (E, D), (E, F), (E, G)
		(F, D), (F, E), (F, G), (F, H)
		(G, D), (G, E), (G, F), (G, H)
		(H, F), (H, G)
(**M**edium, **S**mall)	16	(A, A), (A, B), (A, C)
		(B, A), (B, B), (B, C)
		(C, A), (C, B), (C, C)
		(D, E)
		(E, D), (E, E)
		(F, G)
		(G, F), (G, G)
		(H, H)
(**L**arge, **S**mall)	17	(A, A), (A, F), (A, G)
		(B, B), (B, H)
		(C, D), (C, E)
		(D, C), (D, E)
		(E, C), (E, D)
		(F, A), (F, G)
		(G, A), (G, F), (G, G)
		(H, B)

Table 7: The Binary Constraint Table

For those of you that are interested, this problem has **12** complete and consistent assignments (note that assignment (A, A, A) violates the requirements outlined in the Appendix but is valid based on the unary and binary constraints we defined above). Alright let's get to work and make sure everyone receives their Christmas presents on time!

Variable Tuple	Total	Assignments
(**L**arge, **M**edium, **S**mall)	11	(A, A, A), (A, B, A), (A, C, A)
		(B, A, B), (B, C, B)
		(C, D, E), (C, E, D), (C, E, E)
		(E, A, C), (E, B, C), (E, C, C)
		(G, F, G)

Table 8: Complete and Consistent Assignments

4.1 CSP Q1 [1pt]

Determining what to define as the variables and what to define as the domains for CSP problems is extremely important. If we had defined the ports as the variables in our approach to this CSP problem, we'd have had 8 variables with a search space of 16777216 complete assignments! By defining the ships as the variables, our search space is significantly smaller.

How many **complete assignments** are there in total for this CSP problem? For an assignment to be complete, all variables must be assigned a single value from its starting domain and the assignment does not have to be consistent. (*AIMA 4th Ed. Section 6.1*)

Answer as a whole number.

Answer: $\boxed{512}$

Solution: A complete assignment is one in which every variable is assigned a value (R&N 6.1). Note that this ignores constraints (which is handled by the definition of consistency). Since there are 3 variables and each variable's domain has 8 values that it can be assigned, there are 8^3 complete assignments or 512.

4.2 CSP Q2 [1pt]

In order to reach a solution quickly, we can make optimizations to our algorithm to perform domain reduction in such a way that contradictions can be detected early so that backtracking may occur. One such process is **forward checking**. In forward checking, whenever a variable X is assigned, the forward-checking process establishes arc-consistency for that variable - for each unassigned variable Y that is connected to X by a constraint, we delete from Y's domain any value that is not arc-consistent with the value assigned to X (*AIMA 4th Ed. Section 6.3.2*).

Small: {A, B, C, D, E, F, G, H}	**M**edium: {A, B, C, D, E, F, G, H}	**L**arge: {A, B, C, E, G, H}

Figure 24: Domain State Zero

Given the node consistent domain, Domain State Zero, in Figure 24, how many values are removed in total from the domains of **S**mall and **L**arge if we assign the value **C** to **M**edium?

Answer as a whole number.

Answer: $\boxed{8}$

Solution:

Edge	**Allowed Tuples**
(**L**arge, **M**edium)	(A, C), (B, C), (E, C)
(**M**edium, **S**mall)	(C, A), (C, B), (C, C)

Table 9: Remaining Allowed Tuples

Remaining domain of **L**arge: A, B, E - 3 values removed. Remaining domain of **S**mall: A, B, C - 5 values removed. Altogether 8 values were removed across the domains of **L**arge and **S**mall.

Path Consistency

Small: {A, B, C, D, E, F, G, H}	**M**edium: {A, B, C, D, E, F, G, H}	**L**arge: {A, B, C, E, G, H}

Figure 25: Domain State One

Figure 25, Domain State One, shows the state of domains after applying node consistency and arc consistency to the graph (yes, it's identical to Domain State Zero). As we can see, arc consistency tightens down the unary constraints by using the arc (binary) constraints. But we can do better by using **path consistency** to further tighten the binary constraints by using implicit constraints that are inferred by looking at triples of variables.

A two-variable set $\{X_i, X_j\}$ is path-consistent with respect to a third variable X_m if, for every assignment $\{X_i = a, X_j = b\}$ consistent with the constraints (if any) on $\{X_i, X_j\}$, there is an assignment to X_m that satisfies the constraints on $\{X_i, X_m\}$ and $\{X_m, X_j\}$. The name refers to the overall consistency of the path from X_i to X_j with X_m in the middle. (*AIMA 4th Ed. Section 6.2.3*)

As an example, given the node and arc consistent Domain State One in Figure 25, making **M**edium and **S**mall path consistent with respect to **L**arge results in the removal of **G** and **H** from the domain of **M**edium and some values from the domain of **S**mall.

4.3 CSP Q3 [2pts]

Given the node and arc consistent Domain State One in Figure 25, make **L**arge and **S**mall path consistent with respect to **M**edium. What values are removed from the domain of **L**arge as a result? Select all that apply.

Multiple select.

- ☐ A
- ☐ B
- ☐ C
- ☐ E
- ☐ G
- ☒ H

4.4 CSP Q4 [2pts]

Given the node and arc consistent Domain State One in Figure 25, make **L**arge and **S**mall path consistent with respect to **M**edium. What values are removed from the domain of **S**mall as a result? Select all that apply.

Multiple select.

☐ A

☐ B

☐ C

☐ D

☐ E

☒ F

☐ G

☒ H

Solution for Q3 and Q4:

Variable Tuple (Large, Small)	Medium Satisfying Large	Medium Satisfying Small	Intersection
(A, A)	A, B, C, D, E	A, B, C	A, B, C
(A, F)	A, B, C, D, E	G	None
(A, G)	A, B, C, D, E	F, G	None
(B, B)	A, C, D, E	A, B, C	A, C
(B, H)	A, C, D, E	H	None
(C, D)	A, B, D, E	E	E
(C, E)	A, B, D, E	D, E	D, E
(E, C)	A, B, C, D, F, G	A, B, C	A, B, C
(E, D)	A, B, C, D, F, G	E	None
(F, A)	D, E, G, H	A, B, C	None
(F, G)	D, E, G, H	F, G	G
(G, A)	D, E, F, H	A, B, C	None
(G, F)	D, E, F, H	G	None
(G, G)	D, E, F, H	F, G	F
(H, B)	F, G	A, C	None

Table 10: Complete and Consistent Assignments

The remaining satisfiable (**L**arge, **S**mall) tuples are: (A, A), (B, B), (C, D), (C, E), (E, C), (F, G), (G, G). This leaves the domain of **L**arge as {A, B, C, E, F, G}, removing the value of **H** and leaves the domain of **S**mall as {A, B, C, D, E, G}, removing the value of **F** and **H**.

5 Probability [5pts]

Author: Zhengyu Li

Introduction

Probability and combinatorics are closely related. Combinatorics studies countable sets, while probability uses combinatorics to assign probabilities (values between 0 and 1) to events. Real-world machine learning tasks frequently involve combinatorial structures as we focus on how to model, infer, or predict using graphs, matchings, hierarchies, or other discrete structures behind the data. Moreover, "thinking probabilistically" is a good approach to navigating complex decisions in life, as we want to make decisions that "maximize our expected value".

5.1 Probability Q1 [1pt]

Let X and Y be i.i.d. variables following the normal distribution $\sim \mathcal{N}(0, 1)$. What is the probability that Y is greater than $10X$? Compute the value of $P(Y > 10X)$.

Hint: You do not need to round.

Answer as an exact decimal (no rounding).

Answer: $\boxed{0.5}$

Solution: Since X and Y are mean 0 normal random variables, $Y - 10X$ is also a mean 0 normal random variable. Therefore, $P(Y - 10X > 0) = P(Y > 10X) = 0.5$.

5.2 Probability Q2 [1pt]

Suppose that two integers a and b are uniformly randomly selected from the set $S = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ with replacement. Find the probability that $\max(0, a) = \min(0, b)$.

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p + q$. *Hint: The absolute value of $p - q$ is 56.*

Answer as a whole number.

Answer: $\boxed{106}$

Solution: The solution to the problem is the probability that $a \leq 0$ and $b \geq 0$.

$$\frac{5^2}{9^2} = 25/81$$

$$25 + 81 = 106$$

5.3 Probability Q3 [1pt]

Your friend proposes a game as follows: in each round, a wheel uniformly numbered 1 through 6 is spun four times. If any of the four numbers match, Player 2 gives Player 1 a dollar. If none of the four numbers match, Player 1 gives a dollar to Player 2. What is the expected value for the number of dollars Player 1 will earn in one round?

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p+q$. *Hint: The absolute value of $p-q$ is 5.*

Answer as a whole number.

Answer: $\boxed{13}$

Solution: The probability of getting 0 matches can be calculated as the number of ways we get 4 unique numbers divided by all possible ways we can spin 4 times, which is

$$\frac{6 \times 5 \times 4 \times 3}{6^4} = \frac{5}{18}$$

So the expected payoff is

$$\frac{13}{18} \times 1 - \frac{5}{18} \times 1 = \frac{8}{18} = \frac{4}{9}$$
$$4 + 9 = 13$$

5.4 Probability Q4 [1pt]

Your friend is back with a new game! In each round, a wheel uniformly numbered 1 through 6 is spun repeatedly until the number 4 appears. You win an amount in dollars equal to the highest number that is encountered during the spins. What is the expected value for the amount of dollars you will win from one round of the game?

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p+q$. *Hint: The absolute value of $p-q$ is 25.*

Hint: Think about the order in which the numbers 4, 5, and 6 appear during the spins. What are the probabilities of each number being the highest before the round ends? Consider how permutations of these numbers affect the outcome.

Answer as a whole number.

Answer: $\boxed{37}$

Solution: Let k be the highest number that we are looking for. Since we will eventually land on a 4, the possible highest number must be 4, 5, 6.

- $P(k = 4) = \frac{1}{3}$, since the only case we have this is we see 4 before 5 or 6, and the probability of this is $\frac{1}{3}$.
- $P(k = 5) = \frac{1}{6}$, since the only way this happens is if we see a 5 before seeing a 4, and we never see a 6. Of the 3! arrangements of 4,5,6, only one arrangement works.
- $P(k = 6) = \frac{1}{2}$, since out of 3! arrangements, half of them will see 6 before 4.

$$E[k] = 4 \times \frac{1}{3} + 5 \times \frac{1}{6} + 6 \times \frac{1}{2} = 31/6$$
$$31 + 6 = 37$$

5.5 Probability Q5 [1pt]

Consider a bag of marbles containing red and green marbles where the probability of drawing a red marble from the bag is 0.5 and the probability of drawing a green marble from the bag is also 0.5. Two friends, A and B, each have their own bag of marbles and they continuously draw marbles from their own bags with replacement, stopping when they draw a red marble. What is the probability that person A draws more marbles than person B?

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p + q$. *Hint: The absolute value of $p - q$ is 2.*

Hint: Consider the probability distribution of the number of marbles each friend draws until they get a red marble. Using symmetry in the problem can simplify the calculation.

Answer as a whole number.

Answer: $\boxed{4}$

Solution:

1. **Calculate $P(B < A)$:** Since the events $B < A$ and $A < B$ are symmetric, we have:

$$P(B < A) = \frac{1}{2}(1 - P(A = B))$$

2. **Determine $P(A = B)$:** The probability that both A and B draw their first red marble on the same number of draws follows a geometric sequence. Specifically, there is a probability of $\frac{1}{4}$ that both draw a red marble on their first draw, $\frac{1}{16}$ that both do so on the second draw, and so on. This leads to:

$$P(A = B) = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \cdots$$

   Summing this series gives:

$$P(A = B) = \frac{1}{3}$$

3. **Compute $P(B < A)$:** Using $P(A = B) = \frac{1}{3}$, we find:

$$P(B < A) = \frac{1}{2}\left(1 - \frac{1}{3}\right) = \frac{1}{3}$$

4. $1 + 3 = 4$

6 Bayes Nets [10pts]

Author: Nicolas Shu

Introduction

A lot of the questions on this section are meant to be solved in the same way you solved your implementation of Gibb's sampling in Assignment 3: you are given a problem and you have some probability tables. Your task is to derive a probability formula that equates the conditional probability distribution in terms of the probability tables that you have.

Cancer Progression Network Involving the p53 gene

The p53 gene is a pretty cool gene. It is sometimes referred to as the "guardian of the genome", as it plays a crucial role in preventing cancer and maintaining cellular integrity. Below are a few of its functions:

- It helps prevent the development of tumors by encoding a protein that regulates the cell cycle.
- Whenever there is DNA damage, it can initiate repair processes.
- If the DNA damage is too severe to repair, the p53 gene can halt the cell cycle by inducing cell cycle arrest (i.e. prevents the cell from going through mitosis).
- If the repair is not possible, the gene can induce apoptosis (i.e. programmed cell death), thus eliminating cells that could be cancerous.

The p53 gene can also be activated in response to various types of cellular stress, such as hypoxia and oxidative stress. With this information, we can build a Bayesian network.

The variables in our network are:

1. D: DNA Damage
2. S: Stress Signals
3. M: Mutation in p53 Gene
4. R: DNA Repair
5. A: Apoptosis
6. P: Cell Proliferation
7. C: Cancer Development

The structure for the network is as follows:

- DNA Damage (D) and Stress Signals (S) influence the Mutation in the p53 Gene (M).
- Mutation in p53 (M) and DNA Damage (D) influence DNA Repair (R).
- Mutation in p53 (M) influences Cell Proliferation (P) and Apoptosis (A).
- Ineffective DNA Repair (R), increased Cell Proliferation (P), and failure to trigger Apoptosis (A) influence Cancer Development (C).

Luckily, in molecular biology, we have a number of assays which allows us to check for many of these variables (E.g. the γ H2AX assay can be used to assess DNA damage, and the MTT assay can be used to check for cell proliferation).

Our probability tables are:

1. $p(D)$: Probability of DNA Damage.
2. $p(S)$: Probability of Stress Signals.
3. $p(M|D,S)$: Probability of Mutation in p53 Gene given DNA Damage and Stress Signals.
4. $p(R|D,M)$: Probability of DNA Repair given DNA Damage and Mutation in p53 Gene.
5. $p(A|M)$: Probability of Apoptosis given Mutation in p53 Gene.
6. $p(P|M)$: Probability of Cell Proliferation given Mutation in p53 Gene.
7. $p(C|P,A,R)$: Probability of Cancer Development given Cell Proliferation, Apoptosis, and DNA Repair.

Probability Tables

DNA Damage (D)	$p(D)$
Low	0.6
High	0.4

Stress Signals (S)	$p(S)$
Absent	0.7
Present	0.3

DNA Damage (D)	Stress Signals (S)	Mutation (M)	$p(M|D,S)$
Low	Absent	Absent	0.7
Low	Absent	Present	0.3
Low	Present	Absent	0.4
Low	Present	Present	0.6
High	Absent	Absent	0.5
High	Absent	Present	0.5
High	Present	Absent	0.2
High	Present	Present	0.8

DNA Damage (D)	Mutation (M)	DNA Repair (R)	$p(R|D,M)$
Low	Absent	Effective	0.7
Low	Absent	Ineffective	0.3
Low	Present	Effective	0.6
Low	Present	Ineffective	0.4
High	Absent	Effective	0.5
High	Absent	Ineffective	0.5
High	Present	Effective	0.4
High	Present	Ineffective	0.6

Mutation (M)	Apoptosis (A)	$p(A|M)$
Absent	Not Triggered	0.8
Absent	Triggered	0.2
Present	Not Triggered	0.4
Present	Triggered	0.6

Mutation (M)	Cell Proliferation (P)	$p(P|M)$
Absent	Normal	0.6
Absent	Increased	0.4
Present	Normal	0.3
Present	Increased	0.7

Cell Proliferation (P)	Apoptosis (A)	DNA Repair (R)	Cancer (C)	$p(C\|P, A, R)$
Normal	Not Triggered	Effective	Absent	0.8
Normal	Not Triggered	Effective	Present	0.2
Normal	Not Triggered	Ineffective	Absent	0.6
Normal	Not Triggered	Ineffective	Present	0.4
Normal	Triggered	Effective	Absent	0.4
Normal	Triggered	Effective	Present	0.6
Normal	Triggered	Ineffective	Absent	0.7
Normal	Triggered	Ineffective	Present	0.3
Increased	Not Triggered	Effective	Absent	0.6
Increased	Not Triggered	Effective	Present	0.4
Increased	Not Triggered	Ineffective	Absent	0.8
Increased	Not Triggered	Ineffective	Present	0.2
Increased	Triggered	Effective	Absent	0.3
Increased	Triggered	Effective	Present	0.7
Increased	Triggered	Ineffective	Absent	0.5
Increased	Triggered	Ineffective	Present	0.5

6.1 Bayes Nets Q1 [1pt]

What is the probability of a cell being a cancer cell (C = Present) given that there is no DNA damage (D = Low), there is a mutation in the p53 gene (M = Present), there is a lot of cell proliferation (P = Increased), DNA repair is ineffective (R = Ineffective), apoptosis has not been triggered (A = Not Triggered), and the cell is having stress signals (S = Present)?

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p + q$. *Hint: The absolute value of $p - q$ is 4.*

Answer as a whole number.

Answer: $\boxed{6}$

Solution: We know R, A, and P, and that $C \perp D, M, S$. Therefore,

$$\begin{aligned} p(C|D, M, P, R, A, S) &= p(C|P, A, R) \\ &= p(C = \text{present}|P = \text{increased}, A = \text{not triggered}, R = \text{ineffective}) = \frac{1}{5} \end{aligned}$$

$$1 + 5 = 6$$

6.2 Bayes Nets Q2 [1pt]

In the question above (Q1), C is D-separated from which of the following variables? Select all that apply.

Multiple select.

- ☒ D
- ☒ S
- ☒ M
- ☐ R
- ☐ A

☐ P

Solution: Given that we know the cell proliferation, the DNA repair, and the apoptosis state, and that they form the Markov blanket for C, all other variables are independent as a result.

6.3 Bayes Nets Q3 [2pts]

What is the probability of the cell having a high DNA damage (D = High), given no stress signals (S = Absent) and the mutation of the p53 gene is absent (M = Absent)? For this question only, assume $P(S|D) = P(S)$.

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p + q$. *Hint: The absolute value of $p - q$ is 21.*

Answer as a whole number.

Answer: $\boxed{41}$

Solution: Here we want to know D = High, given S = Absent and M = Absent.

$$\begin{aligned}
p(D|S, M) &= \frac{p(D, S, M)}{p(S, M)} \\
&= \frac{p(S, M|D)p(D)}{p(S, M)} \\
&= \frac{p(M|D, S)p(S|D)p(D)}{p(S, M)} \\
&= \frac{p(M|D, S)p(S)p(D)}{p(S, M)}
\end{aligned}$$

Then, we know that

$$\begin{aligned}
p(S, M) &= \sum_D p(S, M, D) \\
&= \sum_D p(S, M|D)p(D) \\
&= \sum_D p(M|D, S)p(S)p(D)
\end{aligned}$$

Therefore,

$$\begin{aligned}
p(D|S, M) &= \frac{p(M|D, S)p(S)p(D)}{p(S, M)} \\
&= \frac{p(M|D, S)p(S)p(D)}{\sum_D p(M|D, S)p(S)p(D)} \\
&= \frac{p(M|D, S)p(D)}{\sum_D p(M|D, S)p(D)} \\
&= \frac{0.5 \cdot 0.4}{(0.7 \cdot 0.6) + (0.5 \cdot 0.4)} = \frac{10}{31}
\end{aligned}$$

And of course,

$$10 + 31 = 41$$

6.4 Bayes Nets Q4 [3pts]

What is the probability of a cell being cancerous (C), given that the DNA repair is effective ($R =$ Effective), that no apoptosis is occurring ($A =$ Not Triggered), and that there's no mutation on the p53 gene ($M =$ Absent)?

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p+q$. *Hint: The absolute value of* $p-q$ *is 18.*

Answer as a whole number.

Answer: $\boxed{32}$

Solution: Here we know that $C =$ Present, $R =$ Effective, $A =$ Not Triggered, and $M =$ absent.

$$\begin{aligned} p(C|R, A, M) &= \sum_P p(C|P, A, R)p(P|M) \\ &= p(C|P = normal, A, R)p(P = normal|M) + p(C|P = increased, A, R)p(P = increased|M) \\ &= (0.2 \cdot 0.6) + (0.4 \cdot 0.4) = \frac{7}{25} \end{aligned}$$

$$7 + 25 = 32$$

6.5 Bayes Nets Q5 [3pts]

What is the probability of the cell proliferation being increased ($P =$ Increased), given that the apoptosis has been triggered ($A =$ Triggered), and that the DNA repair is ineffective ($R =$ Ineffective)?

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p+q$. *Hint: The absolute value of* $p-q$ *is 1167.*

Answer as a whole number.

Answer: $\boxed{5243}$

Solution: Here we know that $P = \text{Increased}$, $A = \text{Triggered}$, $R = \text{Ineffective}$.

$$
\begin{aligned}
p(P|A,R) &= \frac{p(P,A,R)}{p(A,R)} \\
p(P,A,R) &= \sum_M \sum_D \sum_S p(P|M)p(A|M)p(R|M,D)p(M|D,S)p(D)p(S) \\
= \quad & p(P_{\text{inc}}|M_{\text{abs}})p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{low}}, S_{\text{abs}})p(D_{\text{low}})p(S_{\text{abs}}) \\
& + p(P_{\text{inc}}|M_{\text{pre}})p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{low}}, S_{\text{abs}})p(D_{\text{low}})p(S_{\text{abs}}) \\
& + p(P_{\text{inc}}|M_{\text{abs}})p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{high}}, S_{\text{abs}})p(D_{\text{high}})p(S_{\text{abs}}) \\
& + p(P_{\text{inc}}|M_{\text{pre}})p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{high}}, S_{\text{abs}})p(D_{\text{high}})p(S_{\text{abs}}) \\
& + p(P_{\text{inc}}|M_{\text{abs}})p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{low}}, S_{\text{pre}})p(D_{\text{low}})p(S_{\text{pre}}) \\
& + p(P_{\text{inc}}|M_{\text{pre}})p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{low}}, S_{\text{pre}})p(D_{\text{low}})p(S_{\text{pre}}) \\
& + p(P_{\text{inc}}|M_{\text{abs}})p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{high}}, S_{\text{pre}})p(D_{\text{high}})p(S_{\text{pre}}) \\
& + p(P_{\text{inc}}|M_{\text{pre}})p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{high}}, S_{\text{pre}})p(D_{\text{high}})p(S_{\text{pre}}) \\
= \quad & 0.4 \cdot 0.2 \cdot 0.3 \cdot 0.7 \cdot 0.6 \cdot 0.7 \\
& + 0.7 \cdot 0.6 \cdot 0.4 \cdot 0.3 \cdot 0.6 \cdot 0.7 \\
& + 0.4 \cdot 0.2 \cdot 0.5 \cdot 0.5 \cdot 0.4 \cdot 0.7 \\
& + 0.7 \cdot 0.6 \cdot 0.6 \cdot 0.5 \cdot 0.4 \cdot 0.7 \\
& + 0.4 \cdot 0.2 \cdot 0.3 \cdot 0.4 \cdot 0.6 \cdot 0.3 \\
& + 0.7 \cdot 0.6 \cdot 0.4 \cdot 0.6 \cdot 0.6 \cdot 0.3 \\
& + 0.4 \cdot 0.2 \cdot 0.5 \cdot 0.2 \cdot 0.4 \cdot 0.3 \\
& + 0.7 \cdot 0.6 \cdot 0.6 \cdot 0.8 \cdot 0.4 \cdot 0.3 \\
& = 0.007056 + 0.021168 + 0.0056 + 0.03528 + 0.001728 + 0.018144 + 0.00096 + 0.024192 \\
& = 0.114128 = \frac{7133}{62500}
\end{aligned}
$$

$$
\begin{aligned}
p(A,R) = \sum_M \sum_D \sum_S p(A|M)p(R|M,D)p(M|D,S)p(D)p(S) \\
= \quad & p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{low}}, S_{\text{abs}})p(D_{\text{low}})p(S_{\text{abs}}) \\
& +p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{low}}, S_{\text{abs}})p(D_{\text{low}})p(S_{\text{abs}}) \\
& +p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{high}}, S_{\text{abs}})p(D_{\text{high}})p(S_{\text{abs}}) \\
& +p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{high}}, S_{\text{abs}})p(D_{\text{high}})p(S_{\text{abs}}) \\
& +p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{low}}, S_{\text{pre}})p(D_{\text{low}})p(S_{\text{pre}}) \\
& +p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{low}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{low}}, S_{\text{pre}})p(D_{\text{low}})p(S_{\text{pre}}) \\
& +p(A_{\text{trig}}|M_{\text{abs}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{abs}})p(M_{\text{abs}}|D_{\text{high}}, S_{\text{pre}})p(D_{\text{high}})p(S_{\text{pre}}) \\
& +p(A_{\text{trig}}|M_{\text{pre}})p(R_{\text{inef}}|D_{\text{high}}, M_{\text{pre}})p(M_{\text{pre}}|D_{\text{high}}, S_{\text{pre}})p(D_{\text{high}})p(S_{\text{pre}}) \\
= \quad & 0.2 \cdot 0.3 \cdot 0.7 \cdot 0.6 \cdot 0.7 \\
& +0.6 \cdot 0.4 \cdot 0.3 \cdot 0.6 \cdot 0.7 \\
& +0.2 \cdot 0.5 \cdot 0.5 \cdot 0.4 \cdot 0.7 \\
& +0.6 \cdot 0.6 \cdot 0.5 \cdot 0.4 \cdot 0.7 \\
& +0.2 \cdot 0.3 \cdot 0.4 \cdot 0.6 \cdot 0.3 \\
& +0.6 \cdot 0.4 \cdot 0.6 \cdot 0.6 \cdot 0.3 \\
& +0.2 \cdot 0.5 \cdot 0.2 \cdot 0.4 \cdot 0.3 \\
& +0.6 \cdot 0.6 \cdot 0.8 \cdot 0.4 \cdot 0.3 \\
& = 0.01764 + 0.03024 + 0.014 + 0.0504 + 0.00432 + 0.02592 + 0.0024 + 0.03456 \\
& = 0.17948 = \frac{4487}{25000}
\end{aligned}
$$

$$p(P|A,R) = p(P,A,R)/p(A,R) = \frac{2038}{3205}$$

$$2038 + 3205 = 5243$$

6.6 Bayes Nets Q6 - Extra Credit [0.5pts]

What is the probability of the cell proliferation being increased (P = Increased), given that the apoptosis has been triggered (A = Triggered), that there is a mutation on the p53 gene (M = Present), and that the DNA repair is ineffective (R = Ineffective)?

The answer can be expressed in the form of p/q where p and q are relatively prime positive integers. Submit your answer as the value of $p + q$. *Hint: The absolute value of $p - q$ is 3.*

Answer as a whole number.

Answer: $\boxed{17}$

Solution: Here we know that P = Increased, A = Triggered, R = Ineffective, M = Present.

$$p(P|A,R,M) = p(P|M) = \frac{7}{10}$$

6.7 Bayes Nets Q7 - Extra Credit [0.5pts]

The problems above (Q5 and Q6) generate two very different solutions and inherently serve as an illustrative example of one type of D-separation. Which type of D-separation is that?

Select one.

☐ Chain structure ($X \to Y \to Z$; $X \perp Z|Y$)

☒ Fork structure ($X \leftarrow Y \to Z$; $X \perp Z|Y$)

☐ Collider structure ($X \to Y \leftarrow Z$; $X \not\perp Z|Y$)

Solution: Since Q5 covers $P(P|A, R)$ and Q6 covers $P(P|A, R, M)$, the addition of the variable M on Q6 leads to an independence between A, R which is an example of how the fork structure D-separation independence.

7 Machine Learning [15pts]

Author: Koushik Nagaraj

Gaussian Mixture Models (GMM)

Before starting this section, please refer to the GMM Resource for more information.

A company wants to classify its customers based on their annual income and spending score using a **Gaussian Mixture Model (GMM)**. The dataset contains the following 5 data points. Each data point is given as (X, Y) where $X =$ Income in thousands of dollars and $Y =$ Spending score.

$$\text{Data Points} = \{(10, 30), (12, 35), (15, 40), (18, 45), (20, 50)\}$$

Assume that you want to fit a **GMM with two components (clusters)**. Use the following initial guesses for the means and covariances of the two Gaussian distributions:

- **Component 1:**
  Mean: $\mu_1 = (10, 30)$, Covariance: $\Sigma_1 = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$
- **Component 2:**
  Mean: $\mu_2 = (18, 45)$, Covariance: $\Sigma_2 = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$

The responsibility γ_{kn} of a Gaussian component k for a data point x_n is given by:

$$\gamma_{kn} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{2} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

where:

- π_k is the mixture weight for component k (assume $\pi_1 = \pi_2 = 0.5$).
- $\mathcal{N}(x_n \mid \mu_k, \Sigma_k)$ is the multivariate normal distribution PDF for a 2D Gaussian defined as follows:

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

7.1 Machine Learning Q1 [1pt]

Calculate the responsibility of component 1 (γ_1) for the data point $(15, 40)$. Round your answer to 8 decimal places.

Answer as a decimal rounded to 8 decimal places.

Answer: $\boxed{0.00004998}$

Accepted Answers: Any value in the range 0.00004990 - 0.00005 inclusive.

Solution:

Given Parameters:

- **Data Point**: $\mathbf{x} = \begin{bmatrix} 15 \\ 40 \end{bmatrix}$
- **Component 1**:

  - Mean ($\boldsymbol{\mu}_1$): $\begin{bmatrix} 10 \\ 30 \end{bmatrix}$
  - Covariance (Σ_1): $\begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$
- **Component 2**:
  - Mean ($\boldsymbol{\mu}_2$): $\begin{bmatrix} 18 \\ 45 \end{bmatrix}$
  - Covariance (Σ_2): $\begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$
- **Mixture Weights**: $\pi_1 = \pi_2 = 0.5$

To calculate $\mathcal{N}_1 = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1)$:

Determinant and Inverse of Covariance Matrix:

$$|\Sigma_1| = (5)(5) - (0)(0) = 25$$

$$\Sigma_1^{-1} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{5} \end{bmatrix}$$

Compute $\mathbf{x} - \boldsymbol{\mu}_1$

$$\mathbf{x} - \boldsymbol{\mu}_1 = \begin{bmatrix} 15 \\ 40 \end{bmatrix} - \begin{bmatrix} 10 \\ 30 \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \end{bmatrix}$$

Compute Mahalanobis Distance

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) &= \begin{bmatrix} 5 & 10 \end{bmatrix} \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 5 \\ 10 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ 10 \end{bmatrix} \\ &= (1)(5) + (2)(10) = 5 + 20 = 25\end{aligned}$$

Compute $\mathcal{N}_1$

$$\mathcal{N}_1 = \frac{1}{2\pi\sqrt{25}} \exp\left(-\frac{1}{2} \times 25\right) = \frac{1}{10\pi} \exp(-12.5)$$

Computing the numerical value:

$$\begin{aligned}\mathcal{N}_1 &= \frac{1}{31.4159} \times 3.72665 \times 10^{-6} \\ &\approx 1.18684 \times 10^{-7}\end{aligned}$$

To calculate $\mathcal{N}_2 = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2)$:

Determinant and Inverse of Covariance Matrix:

$$|\Sigma_2| = (8)(8) - (0)(0) = 64$$

$$\Sigma_2^{-1} = \begin{bmatrix} \frac{1}{8} & 0 \\ 0 & \frac{1}{8} \end{bmatrix}$$

Compute $\mathbf{x} - \boldsymbol{\mu}_2$

$$\mathbf{x} - \boldsymbol{\mu}_2 = \begin{bmatrix} 15 \\ 40 \end{bmatrix} - \begin{bmatrix} 18 \\ 45 \end{bmatrix} = \begin{bmatrix} -3 \\ -5 \end{bmatrix}$$

Compute Mahalanobis Distance

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu}_2)^\top \Sigma_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) &= \begin{bmatrix} -3 & -5 \end{bmatrix} \begin{bmatrix} \frac{1}{8} & 0 \\ 0 & \frac{1}{8} \end{bmatrix} \begin{bmatrix} -3 \\ -5 \end{bmatrix} \\
&= \left(\frac{(-3)^2}{8} + \frac{(-5)^2}{8} \right) \\
&= \left(\frac{9}{8} + \frac{25}{8} \right) = \frac{34}{8} = 4.25
\end{aligned}$$

Compute $\mathcal{N}_2$

$$N_2 = \frac{1}{2\pi\sqrt{64}} \exp\left(-\frac{1}{2} \times 4.25 \right) = \frac{1}{16\pi} \exp(-2.125)$$

Computing the numerical value:

$$\begin{aligned}
N_2 &= \frac{1}{50.2655} \times 0.11943 \\
&\approx 0.002375
\end{aligned}$$

Calculating the Responsibility γ_1:

Since $\pi_1 = \pi_2 = 0.5$, the responsibility γ_1 is:

$$\gamma_1 = \frac{\pi_1 \mathcal{N}_1}{\pi_1 \mathcal{N}_1 + \pi_2 \mathcal{N}_2} = \frac{\mathcal{N}_1}{\mathcal{N}_1 + \mathcal{N}_2}$$

Substitute the computed values:

$$\gamma_1 = \frac{1.18684 \times 10^{-7}}{1.18684 \times 10^{-7} + 0.002375}$$

Calculate the denominator:

$$\mathcal{N}_1 + \mathcal{N}_2 = 1.18684 \times 10^{-7} + 0.002375 \approx 0.002375119$$

Compute γ_1:

$$\gamma_1 = \frac{1.18684 \times 10^{-7}}{0.002375119} \approx 4.9975 \times 10^{-5}$$

Rounded to 8 decimal places:

$$\gamma_1 = \boxed{0.00004998}$$

7.2 Machine Learning Q2 [1pt]

Calculate the responsibility of component 2 (γ_2) for the data point $(15, 40)$. Round your answer to 8 decimal places.

Answer as a decimal rounded to 8 decimal places.

Answer: $\boxed{0.99995002}$

Accepted Answers: Any value in the range 0.99995 - 0.99996 inclusive.

Solution: Reusing the values calculated in Q1 we can calcualte the Responsibility γ_2:

Since $\pi_1 = \pi_2 = 0.5$, the responsibility γ_2 is:

$$\gamma_2 = \frac{\pi_2 \mathcal{N}_2}{\pi_1 \mathcal{N}_1 + \pi_2 \mathcal{N}_2} = \frac{\mathcal{N}_2}{\mathcal{N}_1 + \mathcal{N}_2}$$

Substitute the computed values:

$$\gamma_2 = \frac{0.002375}{1.18684 \times 10^{-7} + 0.002375}$$

Calculate the denominator:

$$N_1 + N_2 = 1.18684 \times 10^{-7} + 0.002375 \approx 0.002375119$$

Compute γ_2:

$$\gamma_2 = \frac{0.002375}{0.002375119} \approx 0.99995002$$

Rounded to 8 decimal places:

$$\gamma_2 = \boxed{0.99995002}$$

7.3 Machine Learning Q3 [1pt]

Which component should the data point $(15, 40)$ be assigned to?

Hint: You can use approximate values for exponentials to simplify calculations and ensure responsibilities are easy to interpret.

Select one.

- ☐ Component 1
- ☒ Component 2

Solution:

From previous calculations, we found:

Responsibility of Component 1:

$$\gamma_1 \approx 0.00005$$

Responsibility of Component 2:

$$\gamma_2 \approx 0.99995$$

Since γ_2 is much greater than γ_1, the data point (15, 40) is far more likely to belong to Component 2.

Bayesian Information Criterion (BIC)

Given a dataset with $n = 10000$ data points and various Gaussian Mixture Model (GMM) configurations. The BIC (Bayesian Information Criterion) is calculated using the formula:

$$\text{BIC} = p \cdot \ln(n) - 2\ell$$

Where p is the number of parameters in the model, ℓ is the log-likelihood, and n is the number of data points. Use the natural logarithm (ln, base e). The following table provides the log-likelihoods (ℓ) and the number of parameters (p) for each model with different numbers of components k.

Model	Number of Components (k)	Log-Likelihood (ℓ)	Number of Parameters (p)
A	1	-12000	3
B	2	-11500	6
C	3	-10800	12
D	4	-9500	20
E	5	-9400	35

Table 11: Models

7.4 Machine Learning Q4 [1pt]

What is the BIC for **model A**? Round your answer to 2 decimal places.

Answer as a decimal rounded to 2 decimal places.

Answer: $\boxed{24027.63}$

Solution:

Given:

- Number of data points: $n = 10000$
- Number of parameters: $p = 3$
- Log-likelihood: $\ell = -12000$

$$\ln(10000) = \ln(10^4) = 4\ln(10) \approx 4 \times 2.302585093 \approx 9.210340372$$
$$p \cdot \ln(n) = 3 \times 9.210340372 \approx 27.63102112$$
$$-2\ell = -2 \times (-12000) = 24000$$
$$\text{BIC} = 27.63102112 + 24000 = 24027.63102112 = \boxed{24027.63}$$

7.5 Machine Learning Q5 [1pt]

What is the BIC for **model B**? Round your answer to 2 decimal places.

Answer as a decimal rounded to 2 decimal places.

Answer: $\boxed{23055.26}$

Solution:

$$\ln(10000) = \ln(10^4) = 4\ln(10) \approx 4 \times 2.302585093 \approx 9.210340372$$

$$p \cdot \ln(n) = 6 \times 9.210340372 \approx 55.26204223$$

$$-2\ell = -2 \times (-11500) = 23000$$

$$\text{BIC} = 55.26204223 + 23000 = 23055.26204223$$

$$\text{BIC} = \boxed{23055.26}$$

7.6 Machine Learning Q6 [1pt]

What is the BIC for **model C**? Round your answer to 2 decimal places.

Answer as a decimal rounded to 2 decimal places.

Answer: $\boxed{21710.52}$

Solution:

$$\ln(10000) = \ln(10^4) = 4\ln(10) \approx 4 \times 2.302585093 \approx 9.210340372$$

$$p \cdot \ln(n) = 12 \times 9.210340372 \approx 110.524084464$$

$$-2\ell = -2 \times (-10800) = 21600$$

$$\text{BIC} = 110.524084464 + 21600 = 21710.524084464$$

$$\text{BIC} = \boxed{21710.52}$$

7.7 Machine Learning Q7 [1pt]

What is the BIC for **model D**? Round your answer to 2 decimal places.

Answer as a decimal rounded to 2 decimal places.

Answer: $\boxed{19184.21}$

Solution:

$$\ln(10000) = \ln(10^4) = 4\ln(10) \approx 4 \times 2.302585093 \approx 9.210340372$$

$$p \cdot \ln(n) = 20 \times 9.210340372 \approx 184.20680744$$

$$-2\ell = -2 \times (-9500) = 19000$$

$$\text{BIC} = 184.20680744 + 19000 = 19184.20680744$$

$$\text{BIC} = \boxed{19184.21}$$

7.8 Machine Learning Q8 [1pt]

What is the BIC for **model E**? Round your answer to 2 decimal places.

Answer as a decimal rounded to 2 decimal places.

Answer: $\boxed{19122.36}$

Solution:

$$\ln(10000) = \ln(10^4) = 4\ln(10) \approx 4 \times 2.302585093 \approx 9.210340372$$

$$p \cdot \ln(n) = 35 \times 9.210340372 \approx 322.36191302$$

$$-2\ell = -2 \times (-9400) = 18800$$

$$\text{BIC} = 322.36191302 + 18800 = 19122.36191302$$

$$\text{BIC} = \boxed{19122.36}$$

7.9 Machine Learning Q9 [1pt]

Based on Q4-Q8 what is the model that best fits the data?

Select one.

- ☐ Model A
- ☐ Model B
- ☐ Model C
- ☐ Model D
- ☒ Model E

Solution: Model E best fits the data. The Bayesian Information Criterion (BIC) balances model fit against model complexity. Generally, the model with the lowest BIC value is considered the best fit.

Using the previously computed BIC values:

$$\text{BIC}(A) \approx 24027.63$$
$$\text{BIC}(B) \approx 23055.26$$
$$\text{BIC}(C) \approx 21710.52$$
$$\text{BIC}(D) \approx 19184.21$$
$$\text{BIC}(E) \approx 19122.36$$

Among these, $\text{BIC}(E) \approx 19122.36$ is the smallest. Thus, Model E, which has the lowest BIC, provides the best fit to the data.

Decision Trees

In the evolving domain of marine ecosystem monitoring, real-time fish species classification from underwater drone footage plays a critical role in ocean conservation efforts. This technology aids in identifying fish species to track biodiversity and monitor endangered populations.

The data collected comes from individual video frames captured by underwater drones equipped with advanced sensors. Several important features are extracted from these frames, which serve as input to a decision tree classifier for real-time species identification. Among these features, three are especially critical:

- **Hue-Saturation-Value (HSV)**: A continuous variable representing the color distribution within the frame. HSV is particularly useful in distinguishing fish species based on unique coloration patterns visible in their natural habitat.
- **Water Temperature (WT)**: Another continuous variable, recorded in degrees Celsius, as different species tend to prefer specific temperature ranges. WT helps refine classification by correlating species occurrences with water conditions.
- **Behavioral Pattern (BP)**: A categorical variable describing the activity state of the fish. This feature takes values like 'feeding', 'migrating', or 'resting', providing valuable information for distinguishing between species with similar appearance but different behavioral tendencies.

Using these features, a decision tree classifier divides the input space, enabling rapid species classification for conservationists and marine biologists. This classification empowers data-driven decisions, such as identifying vulnerable populations or environmental changes. See Information Gain Resources for more details.

The dataset for classifying fish species from underwater drone footage can be seen in Table 12 below.

ID	HSV (Color Range)	Water Temperature (°C)	Behavioral Pattern (BP)	Fish Species
1	0.75	24.0	feeding	clownfish
2	0.62	28.5	migrating	tuna
3	0.83	19.0	resting	angelfish
4	0.70	26.0	feeding	clownfish
5	0.58	27.5	migrating	tuna
6	0.85	20.0	resting	angelfish
7	0.77	25.0	feeding	clownfish
8	0.61	29.0	migrating	tuna

Table 12: Fish Classification Dataset

This innovative application highlights the potential of decision trees in ecological research, promoting sustainable ocean resource management by accurately identifying fish species in varying environmental contexts.

7.10 Machine Learning Q10 [2pts]

Given the possible behavioral patterns (feeding, migrating, and resting), calculate the entropy of the Behavioral Pattern feature, $H(BP)$, to determine the uncertainty in the distribution of fish activities in the dataset. Round your answer to 3 decimal places.

Answer as a decimal rounded to 3 decimal places.

Answer: $\boxed{1.562}$

Accepted Answers: Any value in the range 1.561 - 1.562 inclusive.

Solution: Entropy Calculation Using $\log_2$:

Step 1: Formula for Entropy:

The formula for entropy is:

$$H = -\sum p_i \log_2(p_i)$$

where p_i is the proportion of each category in the dataset.

Step 2: Calculate Probabilities:

From the dataset:

- **Feeding**: $p_{\text{feeding}} = \frac{3}{8} = 0.375$
- **Migrating**: $p_{\text{migrating}} = \frac{3}{8} = 0.375$
- **Resting**: $p_{\text{resting}} = \frac{2}{8} = 0.25$

Step 3: Logarithm in Base 2:

Compute the $\log_2$ of each probability:

- $\log_2(0.375) = -1.415$
- $\log_2(0.25) = -2$

Step 4: Weighted Terms:

Multiply each probability by its corresponding $\log_2$ value:

$$p_{\text{feeding}} \log_2(p_{\text{feeding}}) = 0.375 \cdot -1.415 = -0.531$$

$$p_{\text{migrating}} \log_2(p_{\text{migrating}}) = 0.375 \cdot -1.415 = -0.531$$

$$p_{\text{resting}} \log_2(p_{\text{resting}}) = 0.25 \cdot -2 = -0.500$$

Step 5: Sum the Values:

Sum up all the weighted terms:

$$H = -(-0.531 - 0.531 - 0.500)$$

$$H(BP) = 1.562$$

7.11 Machine Learning Q11 [2pts]

We are deciding between two split options for the first node in the decision tree (splitting all of the data). The two potential conditions for splitting our decision tree are as follows:

1. Split A: Split based on the condition HSV > 0.70.
2. Split B: Split based on the condition Water Temperature $< 25.0°C$.

Which split should we use?

Select one.

☒ Split A, because it has the higher Information Gain and is therefore more useful.

☐ Split B, because it has the higher Information Gain and is therefore more useful.

☐ Either split because split A and split B have equal Information Gain and are therefore equivalently useful.

☐ Cannot decide due to lack of information.

Solution:

Step 1: Calculate the initial entropy of the dataset with respect to the target variable (Fish Species).

Dataset species distribution:

$$\begin{aligned} \text{Clownfish: } & 3 \text{ occurrences} \\ \text{Tuna: } & 3 \text{ occurrences} \\ \text{Angelfish: } & 2 \text{ occurrences} \\ |S| & = 8 \end{aligned}$$

Proportions:

$$p(\text{clownfish}) = \frac{3}{8} = 0.375, \quad p(\text{tuna}) = \frac{3}{8} = 0.375, \quad p(\text{angelfish}) = \frac{2}{8} = 0.25$$

Entropy:

$$H(S) = -\sum_i p_i \log_2(p_i) = -(0.375 \log_2(0.375) + 0.375 \log_2(0.375) + 0.25 \log_2(0.25))$$

Calculating logs:

$$\log_2(0.375) \approx -1.415, \quad \log_2(0.25) = \log_2\left(\frac{1}{4}\right) = -2$$

Substitute:

$$H(S) = -[0.375(-1.415) + 0.375(-1.415) + 0.25(-2)]$$

$$H(S) = -[-0.530625 - 0.530625 - 0.5] = -(-1.56125) = 1.56125$$

Initial entropy:

$$H(S) \approx 1.561$$

Step 2: Calculate Information Gain for Split A (HSV $>$ 0.70).

Split A divides the dataset into:

$$A_1 : HSV > 0.70, \quad A_2 : HSV \leq 0.70$$

Group A_1:

$$\text{Instances: clownfish=2, angelfish=2, tuna=0, total } |A_1| = 4$$

Proportions:

$$p(\text{clownfish}) = 0.5, \quad p(\text{angelfish}) = 0.5, \quad p(\text{tuna}) = 0$$

$$H(A_1) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = -(0.5(-1) + 0.5(-1)) = -(-1) = 1.0$$

Group A_2:

$$\text{Instances: clownfish=1, tuna=3, angelfish=0, total } |A_2| = 4$$

Proportions:

$$p(\text{clownfish}) = 0.25, \quad p(\text{tuna}) = 0.75, \quad p(\text{angelfish}) = 0$$

$$H(A_2) = -(0.25\log_2(0.25) + 0.75\log_2(0.75))$$

$$\log_2(0.25) = -2, \quad \log_2(0.75) \approx -0.415$$

$$H(A_2) = -[0.25(-2) + 0.75(-0.415)] = -[-0.5 - 0.31125] = -(-0.81125) = 0.81125 \approx 0.811$$

Weighted entropy after split A:

$$H(S|A) = \frac{|A_1|}{|S|}H(A_1) + \frac{|A_2|}{|S|}H(A_2) = \frac{4}{8}(1.0) + \frac{4}{8}(0.811) = 0.5(1.0) + 0.5(0.811) = 0.5 + 0.4055 = 0.9055 \approx 0.906$$

Information Gain for Split A:

$$IG(A) = H(S) - H(S|A) = 1.56125 - 0.9055 = 0.65575 \approx 0.656$$

Step 3: Calculate Information Gain for Split B (WT < 25.0°C).

Split B divides the dataset into:

$$B_1 : WT < 25.0^\circ C, \quad B_2 : WT \geq 25.0^\circ C$$

Group B_1:

Instances: clownfish=1, angelfish=2, tuna=0, total $|B_1| = 3$

Proportions:

$$p(\text{clownfish}) = \frac{1}{3} \approx 0.3333, \quad p(\text{angelfish}) = \frac{2}{3} \approx 0.6667$$

$$H(B_1) = -(0.3333\log_2(0.3333) + 0.6667\log_2(0.6667))$$

$$\log_2(1/3) \approx -1.585, \quad \log_2(2/3) \approx -0.585$$

$$H(B_1) = -[0.3333(-1.585) + 0.6667(-0.585)] = -[-0.528 - 0.390] = -(-0.918) = 0.918$$

Group B_2:

Instances: clownfish=2, tuna=3, angelfish=0, total $|B_2| = 5$

Proportions:

$$p(\text{clownfish}) = 0.4, \quad p(\text{tuna}) = 0.6$$

$$H(B_2) = -(0.4\log_2(0.4) + 0.6\log_2(0.6))$$

$$\log_2(0.4) \approx -1.3219, \quad \log_2(0.6) \approx -0.7369$$

$$H(B_2) = -[0.4(-1.3219) + 0.6(-0.7369)] = -[-0.5288 - 0.4421] = -(-0.9709) = 0.9709 \approx 0.971$$

Weighted entropy after split B:

$$|B_1| = 3, \quad |B_2| = 5, \quad |S| = 8$$

$$H(S|B) = \frac{3}{8}(0.918) + \frac{5}{8}(0.971) = 0.375(0.918) + 0.625(0.971) = 0.34425 + 0.606875 = 0.951125 \approx 0.951$$

Information Gain for Split B:

$$IG(B) = H(S) - H(S|B) = 1.56125 - 0.951125 = 0.610125 \approx 0.610$$

Step 4: Compare Information Gains

$$IG(A) \approx 0.656, \quad IG(B) \approx 0.610$$

Since:

$$IG(A) > IG(B),$$

we chose the split on HSV (Split A) because it gives a higher Information Gain.

7.12 Machine Learning Q12 [2pts]

What are the key principles to follow when constructing a decision tree to ensure it is best for practical use? Select all that apply.

Multiple select.

☐ Ensure that the tree uses as many features as possible to make each decision.

☒ Choose attributes for splitting that provide the most significant reduction in entropy or Gini Impurity.

☒ Limit the depth of the tree to prevent overfitting and improve generalization to new data.

☒ Implement mechanisms to handle missing values effectively during tree construction.

☐ Design the tree to be as deep as necessary to perfectly classify the training data.

☒ Optimize the tree structure to minimize computational cost during both training and prediction.

Solution:

1. **Choose attributes for splitting that provide the most significant reduction in entropy or Gini Impurity**
   - **Reason:** This ensures that the tree splits data in a way that maximizes information gain or minimizes impurity. These metrics guide the tree to create the most meaningful splits, improving its predictive power and interpretability.
2. **Limit the depth of the tree to prevent overfitting and improve generalization to new data**
   - **Reason:** Unrestricted depth allows the tree to perfectly classify the training data, but this often leads to overfitting, where the tree captures noise rather than general patterns. Limiting depth helps the tree generalize better to unseen data.
3. **Implement mechanisms to handle missing values effectively during tree construction**
   - **Reason:** Real-world datasets often have missing values, and a decision tree must handle them appropriately to remain robust. Strategies like imputing missing values, splitting based on surrogate splits, or using specific algorithms enable the tree to use as much data as possible without introducing bias.
4. **Optimize the tree structure to minimize computational cost during both training and prediction**
   - **Reason:** Efficient decision trees reduce training and inference times, making them more practical for large datasets or real-time applications. This involves balancing the number of nodes, the size of the dataset considered at each split, and the number of features examined.

Why not the others?

- **Ensure that the tree uses as many features as possible to make each decision**
  - **Reason:** Using all features may overcomplicate the model and increase the risk of overfitting. Decision trees aim to identify the most important features for splitting rather than using all available features.
- **Design the tree to be as deep as necessary to perfectly classify the training data**
  - **Reason:** Perfect classification on training data is often a sign of overfitting. The tree should balance depth and accuracy to avoid being overly complex and capturing noise in the training set.

8 Pattern Recognition Through Time [15pts]

Author: Tanmay Chavan

Dynamic Time Warping (DTW)

Professor Oak, a wildlife biologist, is trying to study exotic species of birds in the Himalayas. He installed recording devices in the trees near their habitat and is studying them via recordings of their sounds and songs. Until now, Professor Oak has identified three distinct songs of the birds. These songs are represented using their pitch recorded at short, regular intervals. The songs can be represented as distinct time series:

$$song_a : [21, 23, 22, 25, 26, 22, 19, 16]$$
$$song_b : [22, 20, 17, 15, 18, 20, 15, 17]$$
$$song_c : [23, 24, 25, 27, 26, 22, 20, 18]$$

Professor Oak wants to study these songs and identify the differences and similarities between them. He believes this would help him obtain additional information about the species of the birds. He consults Professor Juniper, a renowned ornithologist, to find a way to do this. She recommends him to explore Dynamic Time Warping. Dynamic Time Warping is a method used to compare two time series. It is more robust than simply computing the euclidean distances between two time series. It can measure the similarity between two time series effectively despite differences in their time or speed. It is often used in speech recognition tasks.

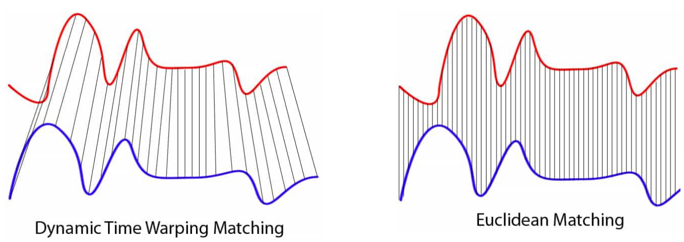


Figure 26: Dynamic Time Warping

Professor Juniper explains how we can implement dynamic time warping by creating a grid and comparing the two songs. For example, consider the following two songs:

$$A = [4, 6, 8, 8, 5, 4, 3, 7]$$
$$B = [3, 4, 6, 9, 8, 5, 2, 6]$$

We can populate each square in the grid using the following formula:

$$D[i, j] = |A_i - B_j| + min(D[i-1, j], D[i, j-1], D[i-1, j-1])$$

She further explains by filling the grid partially:

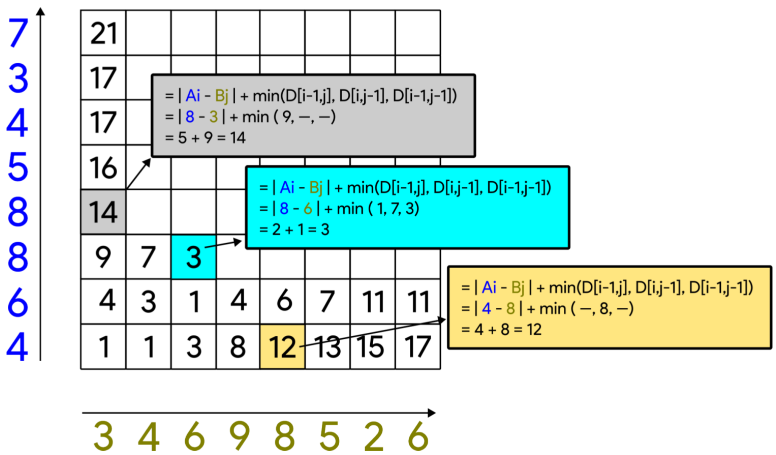


Figure 27: DTW Procedure

After the grid is filled, the distance is given by the lowest path along the matrix, between the top right and the bottom left. You should start from the top right cell and traverse to the bottom left cell. If you start from the top right cell, you can move left, diagonally downwards, or down. In case of a tie, you can move diagonally.

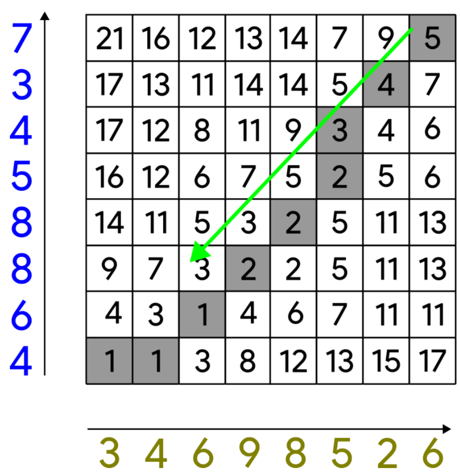


Figure 28: Fully filled grid

For the given two example songs, the shortest path is 1-1-1-2-2-2-3-4-5. **The distance function that we will use in this problem will be to just take the value in the top right cell of the matrix**, which in this case is 5.

Now, you have to help Professor Oak compute the similarities between the songs. For each pair of songs, write down the shortest path and the distance between them. Ensure that while writing the path it starts from the bottom left to the top right and that you separate each value in the cells with hyphens, similar to how we wrote the path in the example above (1-1-1-2-2-2-3-4-5).

Pattern Recognition Through Time Q1 [3pts]

$song_b$								
17			29			39		
15		24			34			
20	16		19			29		18
18				22	**B**		17	**C**
15	12	14				23		12
17			9		16			
20		4		7			10	
22	1				9		**A**	18
	21	23	22	25	26	22	19	16
				$song_a$				

Figure 29: Grid for $song_a$ vs. $song_b$

8.1 Pattern Recognition Through Time Q1a [0.5pts]

What is the value of **A** in Figure 29 above ($song_a$ vs. $song_b$)?

Answer as a whole number.

Answer: $\boxed{12}$

Solution: See Q1e.

8.2 Pattern Recognition Through Time Q1b [0.5pts]

What is the value of **B** in Figure 29 above ($song_a$ vs. $song_b$)?

Answer as a whole number.

Answer: $\boxed{27}$

Solution: See Q1e.

8.3 Pattern Recognition Through Time Q1c [0.5pts]

What is the value of **C** in Figure 29 above ($song_a$ vs. $song_b$)?

Answer as a whole number.

Answer: $\boxed{14}$

Solution: See Q1e.

8.4 Pattern Recognition Through Time Q1d [0.5pts]

What is the distance between $song_a$ and $song_b$ based on Figure 29?

Answer as a whole number.

Answer: $\boxed{20}$

Solution: See Q1e.

8.5 Pattern Recognition Through Time Q1e [1pt]

What is the path in Figure 29?

Answer as a path of numbers (fill in the blanks).

Answer: $\boxed{1\text{-}2\text{-}2\text{-}5\text{-}9\text{-}9\text{-}10\text{-}12\text{-}16\text{-}17\text{-}18\text{-}19\text{-}20}$

Solution:

$song_b$								
17	26	28	29	33	38	39	24	20
15	22	24	25	29	34	35	22	19
20	16	18	19	23	28	29	18	18
18	15	17	18	22	27	27	17	14
15	12	14	15	19	23	23	16	12
17	6	8	9	12	16	16	12	11
20	2	4	4	7	11	11	10	14
22	1	2	2	5	9	9	12	18
$song_a$	21	23	22	25	26	22	19	16

Figure 30: Grid for $song_a$ vs. $song_b$

Pattern Recognition Through Time Q2 [3pts]

18	25			21		11		
20		A			15		6	
22	21		12			5		14
26		12		7			16	
27	15		10		5		B	25
25		5				8		C
24					6			
23	2			5		9		20
$song_c$ / $song_a$	21	23	22	25	26	22	19	16

Figure 31: Grid for $song_a$ vs. $song_c$

8.6 Pattern Recognition Through Time Q2a [0.5pts]

What is the value of **A** in Figure 31 above ($song_a$ vs. $song_c$)?

Answer as a whole number.

Answer: $\boxed{16}$

Solution:

8.7 Pattern Recognition Through Time Q2b [0.5pts]

What is the value of **B** in Figure 31 above ($song_a$ vs. $song_c$)?

Answer as a whole number.

Answer: $\boxed{16}$

Solution:

8.8 Pattern Recognition Through Time Q2c [0.5pts]

What is the value of **C** in Figure 31 above ($song_a$ vs. $song_c$)?

Answer as a whole number.

Answer: $\boxed{22}$

Solution:

8.9 Pattern Recognition Through Time Q2d [0.5pts]

What is the distance between $song_a$ and $song_c$ based on Figure 31?

Answer as a whole number.

Answer: $\boxed{8}$

Solution:

8.10 Pattern Recognition Through Time Q2e [1pt]

What is the path in Figure 31?

Answer as a path of numbers (fill in the blanks).

Answer: $\boxed{2\text{-}2\text{-}4\text{-}4\text{-}5\text{-}5\text{-}5\text{-}6\text{-}8}$

Solution:

$song_c$								
18	25	21	18	21	23	11	7	8
20	22	16	14	15	15	7	6	10
22	21	13	12	10	9	5	8	14
26	20	12	13	7	5	9	16	26
27	15	9	10	6	5	10	16	25
25	9	5	6	4	5	8	14	22
24	5	3	4	4	6	8	13	21
23	2	2	3	5	8	9	13	20
	21	23	22	25	26	22	19	16
	$song_a$							

Figure 32: Grid for $song_a$ vs. $song_c$

Pattern Recognition Through Time Q3 [3pts]

$song_c$								
18	21		16		19		24	
20		15		23				
22				29		35		47
26	15				B		44	
27			18		30			
25		8			C	29		43
24	3		11		24		35	
23	A			18				40
$song_b$	22	20	17	15	18	20	15	17

Figure 33: Grid for $song_b$ vs. $song_c$

8.11 Pattern Recognition Through Time Q3a [0.5pts]

What is the value of **A** in Figure 33 above ($song_b$ vs. $song_c$)?

Answer as a whole number.

Answer: $\boxed{1}$

Solution:

8.12 Pattern Recognition Through Time Q3b [0.5pts]

What is the value of **B** in Figure 33 above ($song_b$ vs. $song_c$)?

Answer as a whole number.

Answer: $\boxed{33}$

Solution:

8.13 Pattern Recognition Through Time Q3c [0.5pts]

What is the value of **C** in Figure 33 above ($song_b$ vs. $song_c$)?

Answer as a whole number.

Answer: $\boxed{26}$

Solution:

8.14 Pattern Recognition Through Time Q3d [0.5pts]

What is the distance between $song_b$ and $song_c$ based on Figure 33?

Answer as a whole number.

Answer: $\boxed{25}$

Solution:

8.15 Pattern Recognition Through Time Q3e [1pt]

What is the path in Figure 33?

Answer as a path of numbers (fill in the blanks).

Answer: $\boxed{\text{1-3-6-11-15-15-15-16-19-19-21-24-25}}$

Solution:

song_c								
18	21	17	16	19	19	21	24	25
20	17	15	18	23	25	25	30	33
22	15	17	22	29	33	35	42	47
26	15	17	22	29	33	36	44	50
27	11	13	18	25	30	33	41	47
25	6	8	13	21	26	29	37	43
24	3	5	11	19	24	27	35	41
23	1	4	10	18	23	26	34	40
	22	20	17	15	18	20	15	17
	song_b							

Figure 34: Grid for $song_b$ vs. $song_c$

8.16 Pattern Recognition Through Time Q4 [1pt]

Professor Oak feels two of these birds are related to each other, while the third is unrelated. He believes we can identify the outlier using the songs. Can you identify which one is the outlier?

Select one.

- ☐ $song_a$
- ☒ $song_b$
- ☐ $song_c$

Solution: The distance between $song_a$ and $song_c$ is the shortest.

Hidden Markov Models (HMMs)

Professor Oak finds out that the outlier bird lets out a specific call when a mountain lion is nearby. Upon further research, he realizes the lion belongs to a species which was considered to be extinct in the area! The professor wants to study the lion's hunting patterns in greater detail. However it is very elusive, evading cameras and humans and only hunting on specific days. Professor Oak arrives at the conclusion that the only way to predict whether the lion is hunting on a particular day or not is by listening to a specific call of the bird, which he refers to as call C.

Professor Oak collects recordings over several weeks and identifies the days when the bird lets out call C, indicating when the lion is out to hunt. Let us denote the occurrence of a call on a particular day by C and its absence by NC. Let us denote the state of the lion being out for hunting by L and its absence by NL.

The professor observes the following sequences of calls spanning over 5 days:

No call, No call, Call, Call, No call

Professor Oak wants to determine on which days the lion was hunting using the above sequence. Professor Oak has studied HMMs, but realizes that computing the probabilities of every possible sequence will be very time consuming. Confused, he visits Professor Juniper for a better approach. She recommends the Viterbi algorithm to him. The Viterbi algorithm uses dynamic programming to compute the most likely sequence, saving a lot of time. Your task is to help Professor Oak find out the most likely hunting sequence using the recording sequence with the help of the Viterbi algorithm.

On the first day, the probability that the lion is out to hunt is 0.5 and the probability that the lion is not out to hunt is also 0.5. The initial state probabilities are $\pi_L = 0.5$, and $\pi_{NL} = 0.5$.

If the lion has hunted on a given day, the probability it will hunt the next day is 0.6 and the probability that it will not hunt the next day is 0.4. If the lion has not hunted on a given day, the probability it will hunt the next day is 0.3 and the probability it will not hunt the next day is 0.7.

The transition probabilities are then given by Table 13:

		Next State	
		L	NL
Current State	L	0.6	0.4
	NL	0.3	0.7

Table 13: Transition Probabilities

If the lion is out to hunt, the probability that it will be observed that the bird lets out the call (C) is 0.85 while the probability that it will be observed that the bird does not let out the call (NC) is 0.15. If the lion is not out to hunt, the probability that it will be observed that the bird lets out the call (C) is 0.15 while the probability that it will be observed that the bird does not let out the call (NC) is 0.85.

The emission probabilities are then given by Table 14:

		Emission Probability	
		L	NL
Observed	C	0.85	0.15
	NC	0.15	0.85

Table 14: Emission Probabilities

To help with your calculations we provide a partially filled trellis in Table 15 below.

	Day 1 NC	Day 2 NC	Day 3 C	Day 4 C	Day 5 NC
L	A	0.019125	C	D	0.0029597754375
NL		B	0.026551875	0.0038689875	

Table 15: Partially filled Trellis

	Day 1 NC	Day 2 NC	Day 3 C	Day 4 C	Day 5 NC
L	0.075	0.019125	0.064483125	0.03288639375	0.0029597754375
NL	0.425	0.252875	0.026551875	0.0038689875	0.011181373875

Table 16: Fully filled Trellis

8.17 Pattern Recognition Through Time Q5 [1pt]

What is the value of A in Table 15? Do NOT round.

Answer as a decimal.

Answer: $\boxed{0.075}$

Solution: $\pi_L * Emission_{L,NC} = 0.5 * 0.85 = 0.075$

8.18 Pattern Recognition Through Time Q6 [1pt]

What is the value of B in Table 15? Do NOT round.

Answer as a decimal.

Answer: $\boxed{0.252875}$

Solution:
$\max(Prev_L * Trans_{L\to NL} * Emit_{NL,NC}\ ,\ Prev_{NL} * Trans_{NL\to NL} * Emit_{NL,NC})$
$= \max(0.0255, 0.252875) = 0.252875$

8.19 Pattern Recognition Through Time Q7 [1pt]

What is the value of C in Table 15? Do NOT round.

Answer as a decimal.

Answer: $\boxed{0.064483125}$

Solution:
$\max(Prev_L * Trans_{L\to L} * Emit_{L,C}\ ,\ Prev_{NL} * Trans_{NL\to L} * Emit_{NL,C})$
$=\max(0.009753750, 0.064483125) = 0.064483125$

8.20 Pattern Recognition Through Time Q8 [1pt]

What is the value of D in Table 15? Do NOT round.

Answer as a decimal.

Answer: $\boxed{0.03288639375}$

Solution:
$\max(Prev_L * Trans_{L\to L} * Emit_{L,C}\ ,\ Prev_{NL} * Trans_{NL\to L} * Emit_{NL,C})$
$= \max(0.03288639375, 0.006770728125) = 0.03288639375$

8.21 Pattern Recognition Through Time Q9 [1pt]

What is the most likely sequence of the lions hunting patterns given the observed bird calling sequence?

Answer as a sequence (fill in the blank).

Answer: $\boxed{NL \to NL \to L \to L \to NL}$

Solution: We choose the most probable state on each day using the trellis.

9 Logic and Planning [15pts]

Author: Zeyu Chang

Canvas Module 9, Ed Lessons Module 9, AIMA Ch. 7 ∼ 10

Part 1

Terminus is an avid space traveler in the year 2424. They want to select a spaceship that can handle multiple roles. The ship they are looking for should have some combination of the following 4 features.

- Large Cargo Space (CS): The ship has a large storage capacity, enabling it to carry significant cargo or supplies for long missions or to transport goods.
- Advanced Weaponry (AW): The ship is equipped with advanced weapons and defense systems, making it suitable for high-risk missions or exploring unknown space.
- Good Living Quarters (LQ): The ship offers spacious, well-equipped living quarters. Suitable for extended journeys.
- High Maintainability (HM): The ship has accessible components and allows efficient in-flight repairs.

Terminus uses a logical circuit, shown in Figure 35 to determine what combinations of features is desirable. Your task is to analyze this circuit and construct a corresponding truth table.

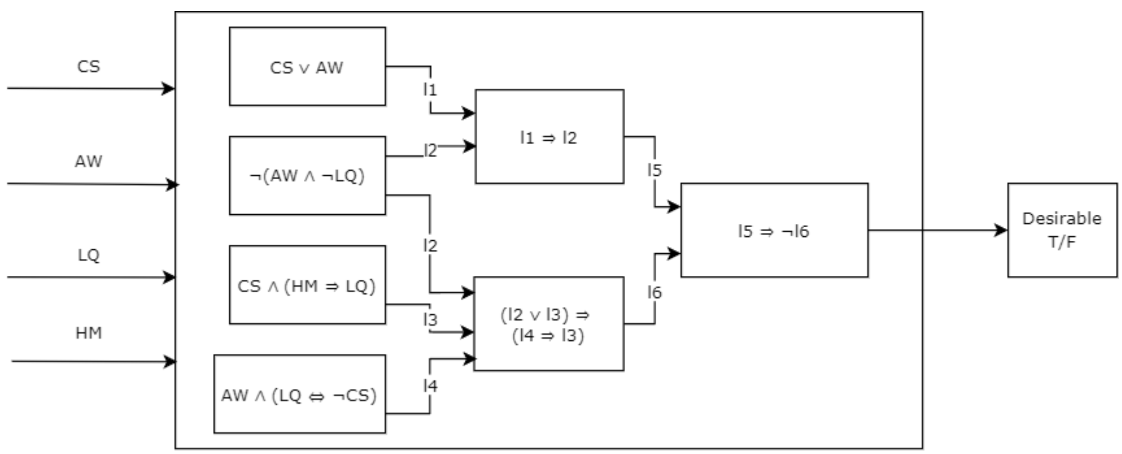


Figure 35: Logical circuit

9.1 Logic and Planning Q1 [4pts]

Given the logical circuit, select all values that are True for **Desirable** for each of the combinations in the partial truth table shown in Table 17 below. Select all that apply.

CS	AW	LQ	HM	Desirable
0	0	0	1	A
0	1	0	0	B
0	1	0	1	C
0	1	1	1	D
1	0	0	0	E
1	0	1	0	F
1	1	1	0	G
1	1	1	1	H

Table 17: Partial Truth Table

Multiple select.

- ☐ A
- ☒ B
- ☒ C
- ☒ D
- ☐ E
- ☐ F
- ☐ G
- ☐ H

Solution: See the attached notebook for the code that implements the circuit. Both a full truth table and a partial truth table is given below.

CS	AW	LQ	HM	Desirable
0	0	0	1	F
0	1	0	0	T
0	1	0	1	T
0	1	1	1	T
1	0	0	0	F
1	0	1	0	F
1	1	1	0	F
1	1	1	1	F

CS	AW	LQ	HM	Desirable
0	0	0	0	F
0	0	0	1	F
0	0	1	0	F
0	0	1	1	F
0	1	0	0	T
0	1	0	1	T
0	1	1	0	T
0	1	1	1	T
1	0	0	0	F
1	0	0	1	F
1	0	1	0	F
1	0	1	1	F
1	1	0	0	T
1	1	0	1	T
1	1	1	0	F
1	1	1	1	F

9.2 Logic and Planning Q2 [2pts]

Terminus is at a spacecraft dealership. They are presented with 2 ships.

- *The Celestial* is a versatile vessel equipped with a large cargo space for hauling extensive supplies across long-distance missions and a high-tech advanced weaponry system for defensive maneuvers when encountering piracy. It lacks a comfortable living quarters, but its high maintainability is hard to miss, as it ensures easy in-flight repairs and upgrades.
- *The Voyager* is designed for space touring, with well-appointed living quarters. However, cargo space is limited and it has outdated weapons. The ship is highly maintainable, but without substantial cargo space or defenses, it may fall short on demanding missions.

According to the truth table, which ship is desirable for Terminus?

Select one.

☒ Only the Celestial is desirable

☐ Only the Voyager is desirable

☐ Both the Celestial and the Voyager are desirable

☐ Neither the Celestial nor the Voyager are desirable

Solution: *The Celestial* corresponds to $CS = 1, AW = 1, LQ = 0, HM = 1$, while *The Voyager* corresponds to $CS = 0, AW = 0, LQ = 1, HM = 1$. According to the truth table, only **Celestial** is desirable.

Part 2

Planet Zyphera has three moons-Orbis, Astraen, and Kaelum. Each moon has one or two cargo facilities, making a total of 5 facilities, and the planet has a cargo center. The **goal** is to transport all the cargo from the facilities to the cargo center (the ship does not have to be at the cargo center). A cargo-hauling ship, which Terminus owns, has a full tank of fuel, and is currently stationed at Zyphera. Their ship can take cargo from 2 facilities at once, but it has bad fuel consumption and must refuel if needed to prevent empty fuel in space. To optimize the mission, the ship's path and actions needed to be planned to avoid unnecessary trips, save fuel, and ensure all cargo reaches the center in minimal time.

Note: Assume that F_x denotes facility x and C_x denotes the cargo from facility x. Assume P denotes Zyphera and also the cargo center. Finally, assume that apart from the cargo from the 5 facilities, there is no other cargo.

Possible actions are:

- *Gather*(c, x): Gather cargo c from location x
- *Move*(x, y): Move ship from x to y, where x, y can be the planet or the location of the facilities. The ship travels fast at the cost of a substantial amount of space-fuel.
- *Offload*(c, x): Offload cargo c at location x.
- *Refuel*(x): Refuel the ship at x.

The predicates are:

- *Cargo*(c) means the cargo c needs to be picked up, while $\neg$*Cargo*(c) means the opposite. If a cargo is delivered to the cargo center, it won't be picked up again.
- *Center*(c) means that the cargo c is at the cargo center (on planet Zyphera).

- *Fuel(F), Fuel(L), Fuel(E)* means the ship's fuel level is full/high (above 30%), low (under 30%), or empty (under 1%), respectively.
- *At(x)* means the ship is at location x.
- *HasCargo(c)* means cargo c is on ship.

The constraints are:

- The ship can only offload cargo when the corresponding cargo is on ship.
- Refueling can only be done at the planet (Zyphera).
- Cargo can only be gathered from a facility when the ship has space for it.
- A move from x to x is invalid.
- To gather or offload at place x, the ship must be at x.

Note: For simplicity, assume that *Cargo(C)* is equivalent to:

$$Cargo(C_1) \wedge Cargo(C_2) \wedge Cargo(C_3) \wedge Cargo(C_4) \wedge Cargo(C_5),$$

and $\neg$*Cargo(C)* means:

$$\neg Cargo(C_1) \wedge \neg Cargo(C_2) \wedge \neg Cargo(C_3) \wedge \neg Cargo(C_4) \wedge \neg Cargo(C_5),$$

similar simplification applies to *Center(·), HasCargo(·)*.

9.3 Logic and Planning Q3 [2pts]

Which of the following clauses represent the initial state of the setup?

Select one.

☐ $Cargo(C)$

☐ $Cargo(C) \wedge Fuel(F)$

☐ $Cargo(C) \wedge Fuel(F) \wedge At(P)$

☐ $Cargo(C) \wedge \neg HasCargo(C) \wedge Fuel(F) \wedge At(P)$

☒ None of the above.

Solution: The initial state should also state that the cargo center does not have the cargoes. Therefore, the correct representation should be:

$$Cargo(C) \wedge \neg HasCargo(C) \wedge \neg Center(C) \wedge Fuel(F) \wedge At(P)$$

9.4 Logic and Planning Q4 [3pts]

Which of the following predicates **must** be in the clause that represent the goal state? Select all that apply.

Multiple select.

☒ $\neg$*Cargo*(C_1)

☐ *HasCargo*(C_2)

☐ *Fuel*(L)

☒ $\neg$*HasCargo*(C_3)

☐ *At*(P)

☒ *Center*(C_4)

☐ $\neg$*Fuel*(F)

Solution: The final state is defined by the event that all cargoes are at the cargo center. There is no requirement on the fuel level or where the ship is.

- $\neg$*Cargo*(C_1) **must be in the clause.** All cargoes should have been picked up by the ship and sent to the cargo center.
- *HasCargo*(C_2) must not be in the clause, as all the cargoes should be delivered to the cargo center and the ship should carry no cargoes.
- *Fuel*(L) need not be in the clause. Fuel level is not defined as part of the final state.
- $\neg$*HasCargo*(C_3) **must be in the clause.**
- *At*(P) need not be in the clause. The ship can be at anywhere as it is not defined as part of the final state.
- *Center*(C_4) **must be in the clause.**
- $\neg$*Fuel*(F) need not be in the clause.

9.5 Logic and Planning Q5 [4pts]

Suppose the ship is currently at F_3. It has delivered cargo C_1, C_5, and has cargo C_2 on board. Cargo C_3, C_4 are still waiting for pick up at F_3, F_4, and the ship's fuel level is low.

If we perform one step of progression search, which of the following are valid next actions? Select all that apply.

Multiple select.

☐ *Offload*(C_3, F_3)

☒ *Move*(F_3, F_1)

☒ *Move*(F_3, P)

☒ *Move*(F_3, F_4)

☐ *Refuel*(F_3)

☒ *Gather*(C_3, F_3)

☒ *Offload*(C_2, F_3)

Solution: Note that the ship is free to move, free to offload or pick up cargo. There is no assumption that the ship will not return to previously visited facilities. Because of this, the only **invalid** next actions in the selection are *Refuel*(F_3) and *Offload*(C_3, F_3). The rest are all possible next actions, even though some of them are clearly not optimal.

10 Planning Under Uncertainty [15pts]

Author: Kai Chen

Introduction

Leela operates an interplanetary delivery service operating between Earth and Mars. Leela's delivery spaceship has limited cargo capacity and needs to decide which packages to accept for transport to maximize profit.

There are three types of packages:

1. Small, high-priority packages with a fixed profit of $3,000 each. They take up 1 unit of cargo space.
2. Medium-sized packages with variable profit. There's a 60% chance of earning $8,000 and a 40% chance of earning $4,000 per package. They take up 2 units of cargo space.
3. Large, mysterious packages. The profit is unknown but rumored to be very high or potentially a loss. They take up 3 units of cargo space.

Leela's spacecraft has a cargo capacity of 5 units. Leela has 3 decision points representing opportunities to accept or reject packages before the spacecraft departs for Earth. At each decision point, Leela can receive a request for one package of any type. Leela would like to maximize the total expected profit from delivered packages.

10.1 Planning Under Uncertainty Q1 [1pt]

What is included in the state space? Select all that apply.

Multiple select.

☒ Remaining decision points

☐ Probability of accepting Package A

☐ Probability of accepting Package B

☒ Available cargo space

Solution: Leela's state space can be described knowing the remaining decision points and available cargo space. Probability of acceptance for packages A, and B is not defined nor part of the state space.

10.2 Planning Under Uncertainty Q2 [1pt]

What is included in the action space? Select all that apply.

Multiple select.

☒ Accept Package (if enough cargo space)

☒ Reject Package

☐ Transition probabilities to next state

☐ Available cargo space

Solution: Actions available to Leela include accepting and rejecting packages (A/B) at each decision point. Transition probabilities are not defined, and available cargo space describes the state space.

10.3 Planning Under Uncertainty Q3 [1pt]

What is the reward for accepting Package A?

Answer as a whole number.

Answer: $\boxed{3000}$

Solution: Directly given in description of package A

10.4 Planning Under Uncertainty Q4 [1pt]

What is the reward for accepting Package B?

Answer as a whole number.

Answer: $\boxed{6400}$

Solution: Computed with $Earning_1 * P(Earning_1) + Earning_2 * P(Earning_2)$ = \$8000*0.6 + \$4000*0.4 = \$6400

10.5 Planning Under Uncertainty Q5 [1pt]

Which of the following statements are true if we only consider packages A and B (assuming package C is not an option)? Select all that apply.

Multiple select.

- ☒ A terminal state is reached when no decision points remain.
- ☒ A terminal state is reached when no cargo space remains.

Solution: Leela cannot accept anymore packages when there are no more decision points or any cargo space remaining.

10.6 Planning Under Uncertainty Q6 [1pt]

Suppose Leela's cargo consist of 2 medium packages and 1 small package. What is Leela's expected profit?

Answer as a whole number.

Answer: $\boxed{15800}$

Solution: We combine the rewards of 2 medium packages and 1 small package, \$6400*2 + \$3000 = \$15800

10.7 Planning Under Uncertainty Q7 [1pt]

Is the policy of carrying 2 medium packages and 1 small package optimal?
Select one.

- ☒ Yes
- ☐ No

Solution: This policy is optimal, a medium package's expected reward is greater than 2 small rewards while taking only 2 cargo spaces.

10.8 Planning Under Uncertainty Q8 [2pts]

Which of the following statements are true? Select all that apply.

Multiple select.

☒ Accepting Package C allows Leela to learn about its profit potential.

☐ Rejecting Package C allows Leela to learn about its profit potential.

☐ Accepting Package C allows Leela to maximize short term gain but potentially miss out on long term gain.

☒ Rejecting Package C lets Leela stick to known profits of packages A and B, maximizing short term gain but potentially missing out on long term gain.

Solution:

- Leela will only gain new information about Package C when she accepts Package C.
- When Leela rejects package C, she either accepts Package A, B, or None (unlikely). Package A and B have defined expected rewards. When accepting package A or B, Leela may potentially miss out on larger gains that can be had when accepting C.

10.9 Planning Under Uncertainty Q9 [3pts]

What methods can be applied to assist when deciding whether to pick package C? Select all that apply.

Multiple select.

☒ Thompson sampling

☒ UCB (Upper confidence bound) heuristic

☒ Gittins Index

☐ Dynamic Time Warping

Solution: Correct selections are for approximating policies to handle the exploration-exploitation dilemma. See Section 16.3.3 - Approximating optimal bandit policies. Dynamic time warping is unrelated to determining optimal policy.

10.10 Planning Under Uncertainty Q10 [3pts]

When deciding whether to select package C, what should Leela take into consideration? Select all that apply.

Multiple select.

☒ Leela's risk tolerance

☒ Number of decision points

☐ Leela's gut feeling

☒ Time horizon

Solution:

- Leela's risk tolerance - Suppose Leela was a big risk taker – Leela will lean more towards exploring potentially gaining large rewards while forgoing known rewards.

- Number of decision points and Time horizon - It may make more sense for Leela to choose Package C given more opportunities to explore (gamble)
- Leela's gut feeling - Although Leela's gut feeling may have historically served Leela well, it is undefinable.

11 Machine Learning - Extra Credit [2pts]

Author: Koushik Nagaraj

All questions in this section are extra credit.

Advancing Agricultural Research with Convolutional Neural Networks (CNN) for Orange Variety Identification

Dr. Muzan has launched a groundbreaking project on classifying orange varieties to support agricultural practices and improve crop quality. The dataset consists of 50,000 high-resolution images of six distinct orange varieties: Valencia, Mandarin, Dream Navel, Kumquat, Torocco, and Chinotto as shown in Figure 36. These varieties exhibit subtle differences in size, color, texture, shape, and rind patterns, which makes precise classification challenging. Accurate identification is essential for inventory management, crop quality monitoring, and market optimization.

For theoretical insights, Dr. Muzan refers to Chapter 21 of Artificial Intelligence: A Modern Approach by Stuart Russell and Peter Norvig, which discusses neural networks and their applications in detail.

Initially, Dr. Muzan opted for a relatively straightforward approach by designing a multi-layer perceptron (MLP) with 6 outputs. His MLP model includes:

Input Image (64 x 64 x 1)
↓
Linear Layer 1 (425 hidden units, bias term, activation function: ReLu)
↓
Linear Layer 2 (255 hidden units, bias term, activation function: ReLu)
↓
Linear Layer 3 (80 hidden units, bias term, activation function: sigmoid)
↓
Output Layer (6 units, bias term, activation function: SoftMax)
↓
output (6 probabilities)

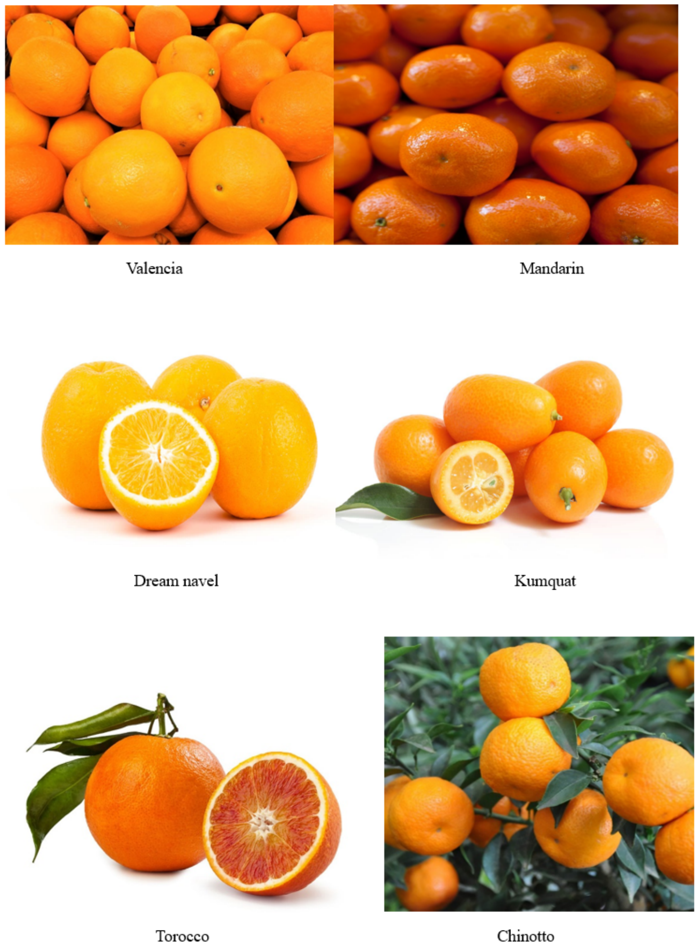


Figure 36: Oranges Dataset

However, when he trains this MLP, it seems to learn very slowly; the training loss doesn't seem to decrease at all, regardless of his optimization procedure. The cause he thinks can be due to the image size or the number of parameters the model holds.

11.1 Machine Learning Extra Credit Q1 [0.4pts]

To answer this, he'd like to know: how many learnable parameters are in this network?

Hint: Remember to include the weights and biases of both layers!

Answer as a whole number.

Answer: $\boxed{1870821}$

Solution: Parameter Calculation Method

For a linear layer with:

$$\text{Input dimension} = I, \quad \text{Output dimension} = O$$

The number of parameters (weights + biases) is:

$$(I \times O) + O$$

This accounts for $I \times O$ weights and O biases.

Step-by-Step Computation

1. **Linear Layer 1:**

$$I_1 = 4096, \quad O_1 = 425$$

$$\text{Parameters}_{L1} = (4096 \times 425) + 425$$

Calculate weights:

$$4096 \times 425 = 4096 \times (400 + 25) = (4096 \times 400) + (4096 \times 25)$$

$$4096 \times 400 = 1{,}638{,}400$$

$$4096 \times 25 = 102{,}400$$

Sum weights:

$$1{,}638{,}400 + 102{,}400 = 1{,}740{,}800$$

Add biases:

$$1{,}740{,}800 + 425 = 1{,}741{,}225$$

2. **Linear Layer 2:**

$$I_2 = 425, \quad O_2 = 255$$

$$\text{Parameters}_{L2} = (425 \times 255) + 255$$

Calculate weights:

$$425 \times 255 = 108{,}375$$

Add biases:

$$108{,}375 + 255 = 108{,}630$$

3. **Linear Layer 3:**

$$I_3 = 255, \quad O_3 = 80$$

$$\text{Parameters}_{L3} = (255 \times 80) + 80$$

Calculate weights:

$$255 \times 80 = 20{,}400$$

Add biases:

$$20{,}400 + 80 = 20{,}480$$

4. **Output Layer:**

$$I_4 = 80, \quad O_4 = 6$$

$$\text{Parameters}_{\text{Output}} = (80 \times 6) + 6$$

Calculate weights:

$$80 \times 6 = 480$$

Add biases:

$$480 + 6 = 486$$

Total Parameters

Sum all parameters:

$$\text{Total} = \text{Parameters}_{L1} + \text{Parameters}_{L2} + \text{Parameters}_{L3} + \text{Parameters}_{\text{Output}}$$
$$= 1{,}741{,}225 + 108{,}630 + 20{,}480 + 486$$

First add $L1 + L2$:

$$1{,}741{,}225 + 108{,}630 = 1{,}849{,}855$$

Add $L3$:

$$1{,}849{,}855 + 20{,}480 = 1{,}870{,}335$$

Add Output layer:

$$1{,}870{,}335 + 486 = 1{,}870{,}821$$

$$\boxed{1{,}870{,}821 \text{ parameters}}$$

Thus, the total number of learnable parameters in the network is **1,870,821**.

11.2 Machine Learning Extra Credit Q2 [0.4pts]

Despite efforts to enhance the multi-layer perceptron (MLP) model's robustness and the allocation of additional computational resources, the model's performance remained suboptimal. This was clearly illustrated in the Loss vs. Epoch/Training size graph, which revealed a concerning trend: despite various attempts at optimization and increasing the model's complexity, there was no significant improvement in the test accuracy.

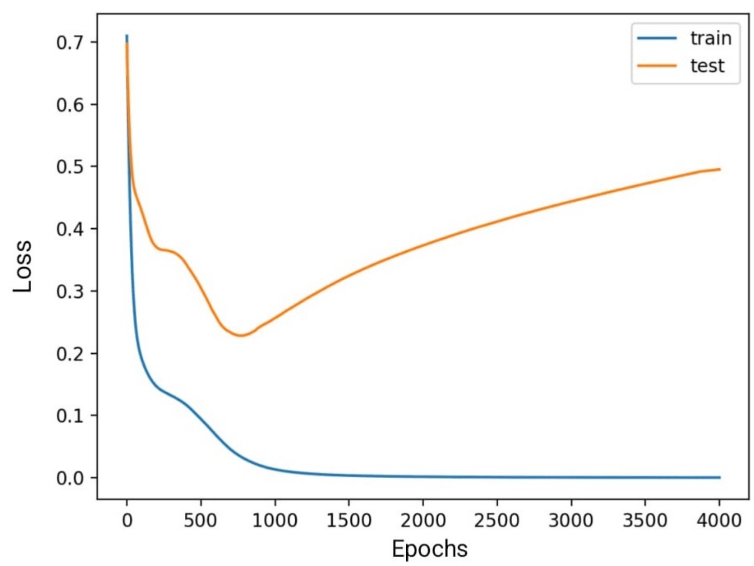


Figure 37: Model Loss vs Epoch/Training Size

What concept best explains the term (problem) observed in Figure 37 above?

Select one.

☐ the curse of dimensionality

☐ underfitting

☒ overfitting

☐ the exploding gradient

Solution: Explaining Overfitting

In the context of machine learning, **overfitting** occurs when a model learns not only the underlying patterns in the training data, but also the noise and random fluctuations. As a result, while the model may perform extremely well on the training dataset, it fails to generalize to new, unseen data.
Formally, consider a training set:

$$\{(x_i, y_i)\}_{i=1}^{N}$$

where x_i represents the input features and y_i the corresponding target values.
An overfit model can be characterized by:

- Exhibiting exceptionally low training error (the model fits every detail of the training set).
- Showing significantly higher error when evaluated on a separate test set or validation set.

In essence, the model is capturing patterns that are specific to the training data but not representative of the general data distribution. This poor generalization performance is the hallmark of overfitting.

11.3 Machine Learning Extra Credit Q3 [0.4pts]

Upon analyzing the multi-layer perceptron (MLP) model with a large number of learnable parameters, Dr.Muzan realizes that while the model has a significant capacity to learn, the sheer number of parameters poses challenges in terms of computational efficiency and potential overfitting, especially given the complexity of the task at hand—classifying orange images.

The MLP's structure, primarily designed for vectorized input data, does not inherently leverage the spatial and hierarchical patterns present in images, which are crucial for effective image classification. Given the complexity of distinguishing these varieties, Dr. Muzan utilizes convolutional neural networks (CNNs) for the task. The transition to a CNN involves structuring layers that can automatically and efficiently recognize patterns in the orange images, such as edges, shapes, and textures, which are essential for distinguishing between different orange types. By employing convolutional layers, we can reduce the number of parameters without sacrificing the network's ability to learn complex patterns. This is achieved by using filters that scan through the image in strides, pooling layers that summarize the features, and ultimately fewer fully connected layers than in our original MLP model. Dr. Muzan then designed a Convolution Neural Network for the task at hand with the architecture below.

- CONV-K-N = convolutional layer with N filters, each size K×K.
- Padding and stride params are consistently set to 0 and 1, respectively.
- POOL-K = K × K pooling layer with a stride K and zero padding.

Recall that there is one bias parameter is added for each filter. For example, in the first convolutional layer (CONV-5-32): there are 32 biases (one per filter).

Input Image (128 x 128 x 1)
↓
Convolution layer(2D) (CONV-5-32)
↓
Max Pooling (2D) (POOL-2)
↓
Convolution layer(2D) (CONV-3-16)
↓
Max Pooling (2D) (POOL-2)
↓

$$\text{Linear Layer 3 (50 hidden units, bias term, activation function: sigmoid)}$$
$$\downarrow$$
$$\text{Output Layer (6 units, bias term, activation function: SoftMax)}$$
$$\downarrow$$
$$\text{output (6 probabilities)}$$

To compare the parameters from a basic feed-forward network and a Convolution Neural Network, Dr. Muzan would like to know how many learnable parameters there are in the above CNN architecture.

Answer as a whole number.

Answer: $\boxed{725812}$

Solution: Calculating the Total Number of Learnable Parameters in the CNN

Given Architecture:

$$\text{Input: } 128 \times 128 \times 1$$

$$\begin{aligned}
&\text{CONV-5-32}: \text{ 32 filters, each of size } 5 \times 5 \times 1 \\
&\text{POOL-2}: 2 \times 2 \text{ max pooling} \\
&\text{CONV-3-16}: \text{ 16 filters, each of size } 3 \times 3 \times 32 \\
&\text{POOL-2}: 2 \times 2 \text{ max pooling} \\
&\text{Linear Layer 3 (Fully Connected): 50 hidden units, bias included, sigmoid activation} \\
&\text{Output Layer: 6 units, bias included, SoftMax activation}
\end{aligned}$$

Dimension Computations and Parameters

Layer 1: CONV-5-32 Input: $128 \times 128 \times 1$ Filter size: $5 \times 5 \times 1$, Number of filters $= 32$ Each filter has $5 \times 5 \times 1 = 25$ weights $+ 1$ bias $= 26$ parameters per filter Total parameters in Layer 1:

$$32 \text{ filters} \times 26 \text{ parameters/filter} = 832$$

After CONV-5-32 (stride 1, no padding): Output dimension:

$$(\text{Width, Height}) = 128 - 5 + 1 = 124$$

So output is $124 \times 124 \times 32$.

Layer 2: POOL-2 Pooling of 2×2 reduces each spatial dimension by factor 2:

$$124 \times 124 \xrightarrow{\text{POOL-2}} 62 \times 62$$

Output after POOL-2: $62 \times 62 \times 32$

Layer 3: CONV-3-16 Input: $62 \times 62 \times 32$ Filter size: $3 \times 3 \times 32$ Number of filters $= 16$
Each filter has:

$$3 \times 3 \times 32 = 288 \text{ weights} + 1 \text{ bias} = 289 \text{ parameters per filter}$$

Total parameters in Layer 3:

$$16 \times 289 = 4624$$

Output dimension after CONV-3-16:

$$62 - 3 + 1 = 60$$

So output is $60 \times 60 \times 16$.

Layer 4: POOL-2 Again, pooling by 2:

$$60 \times 60 \xrightarrow{\text{POOL-2}} 30 \times 30$$

Output after this POOL-2: $30 \times 30 \times 16$

Flattening Before the Fully Connected Layer Flatten $30 \times 30 \times 16$:

$$30 \times 30 \times 16 = 900 \times 16 = 14400 \text{ units}$$

Layer 5: Linear Layer (50 hidden units) Input: 14400 units, Output: 50 units Parameters:

$$(14400 \times 50) \text{ weights} + 50 \text{ biases} = 720000 + 50 = 720050$$

Output Layer: (6 units) Input: 50 units, Output: 6 units Parameters:

$$(50 \times 6) \text{ weights} + 6 \text{ biases} = 300 + 6 = 306$$

Total Parameters

Summing all parameters from each layer:

$$\begin{aligned} \text{CONV-5-32} &: 832 \\ \text{CONV-3-16} &: 4624 \\ \text{Linear Layer (50 units)} &: 720050 \\ \text{Output Layer (6 units)} &: 306 \end{aligned}$$

Add them all:

$$832 + 4624 = 5456$$

$$5456 + 720050 = 725506$$

$$725506 + 306 = 725812$$

$$\boxed{725812 \text{ parameters}}$$

Thus, the total number of learnable parameters in the CNN is **725,812**.

11.4 Machine Learning Extra Credit Q4 [0.4pts]

Dr. Muzan is testing different optimization algorithms to improve the training efficiency of his CNN. During experiments, he notices that with an initial learning rate of 0.01, one optimizer adapts its learning rate over time, resulting in faster convergence compared to a fixed-rate optimizer. Which of the following optimizers is known for having this adaptive learning rate capability?

Select one.

- ☐ Stochastic Gradient Descent (SGD)
- ☒ Adam Optimizer
- ☐ Momentum Optimizer
- ☐ Nesterov Accelerated Gradient (NAG)

Solution: **Answer:** Adam Optimizer

Explanation

Among the listed optimizers:

- **Stochastic Gradient Descent (SGD)** uses a fixed learning rate unless manually adjusted.
- **Momentum Optimizer** and **Nesterov Accelerated Gradient (NAG)** incorporate momentum terms but still require a fixed or manually scheduled learning rate.
- **Adam Optimizer**, on the other hand, combines ideas from both RMSProp and Momentum. It adapts the learning rate for each parameter by estimating first and second moments of the gradients. This leads to an adaptive learning rate mechanism that updates over time without manual intervention.

Therefore, the optimizer known for having this adaptive learning rate capability is the $\boxed{\text{Adam Optimizer}}$.

11.5 Machine Learning Extra Credit Q5 [0.4pts]

Due to the limited availability of data points in the dataset, more data is required to effectively train a robust Convolutional Neural Network (CNN). Dr. Muzan adopted synthetic data generation as a practical and scalable solution to address this challenge. This approach involves generating realistic, computer-created images that closely resemble various types of oranges under different conditions. By incorporating synthetic images that simulate a wide range of features—including orange textures, shapes, sizes, and even surface blemishes—he aimed to improve the network's ability to generalize across a broader array of real-world scenarios.

By enriching the dataset with these diverse synthetic samples, Dr. Muzan sought to mitigate the risks of overfitting and improve CNN's robustness. This enhanced variety ensured that the network would not only excel at recognizing common cases but also perform well in edge cases and novel situations—leading to more reliable, generalizable predictions in real-world applications.

What is one of the primary risks of over-relying on synthetic data for machine learning model development? (Synthetic Dataset Resource for more information).

Select one.

- ☒ The model may learn to perform well on synthetic patterns but fail to generalize effectively to real-world environments.
- ☐ Synthetic data may introduce too much randomness, making it difficult for the model to identify meaningful patterns.
- ☐ The use of synthetic data can eliminate the need for validation using real-world data, reducing deployment timelines.
- ☐ Models trained on synthetic data tend to require more storage space than those trained on real-world datasets.

Solution: The model may learn to perform well on synthetic patterns but fail to generalize effectively to real-world environments.

When a machine learning model is predominantly trained on synthetic data, one primary risk is that it may overfit to features and patterns that are specific to the generated samples. While the model can achieve strong performance on these synthetic inputs, these patterns might not accurately reflect the complexity and nuances of the real world.

As a result, when presented with authentic, unseen data from real-world conditions, such a model may not generalize well. It might struggle to interpret genuine variations that were not sufficiently captured or represented by the synthetic dataset, ultimately leading to poor real-world performance.

Additional Reading

CNNs are well-suited for this project because they can automatically learn hierarchical features. The early layers of the CNN detect low-level patterns (e.g. texture and color), while deeper layers capture more complex structures (e.g. the overall shape or surface irregularities of each orange variety). The use of convolutional layers ensures that the model focuses on local regions of the image to detect essential features, while pooling layers reduce the computational complexity, allowing the system to scale efficiently to large datasets.

Incorporating continuous learning capabilities, the CNN model can adapt to changing growing conditions, harvest timing, or ripening stages. This ensures that the system remains accurate even when environmental factors impact the appearance of the oranges. The ability to monitor crop quality and identify defects early helps farmers and distributors optimize yield and ensures consistent product quality.

Dr. Muzan's decision to use CNNs instead of traditional neural networks is driven by their superior performance with image data. The project not only benefits farmers and agricultural industries but also supports sustainable farming practices by identifying variety-specific needs, such as fertilizer requirements or optimal harvest times.

For more information, please refer to Stuart Russell and Peter Norvig, Artificial Intelligence, A Modern Approach, Chapter 21.3 - (Convolution Neural Network Resource). For more in-depth knowledge, have a look at Masked R-CNN for orange detection (optional resource not mandatory for this exam). This architecture significantly reduces the number of learnable parameters by sharing weights across the spatial dimensions and focusing on local features, making it more computationally efficient and more adept at capturing the nuances of image data.

In light of the challenges encountered with the multi-layer perceptron (MLP)/CNN model, Dr. Muzan turns his attention towards exploring the capabilities of Transformer architecture Recent advancements in using attention mechanisms, like those found in Transformers (originally designed for NLP tasks), have been adapted for image classification (Vision transformer). They offer the potential to focus on specific parts of an image that are more informative for classification, which can be beneficial for distinguishing subtle differences in orange types.

12 Appendix

Constraint Satisfaction Problems (CSPs) - Optional Extra Reading

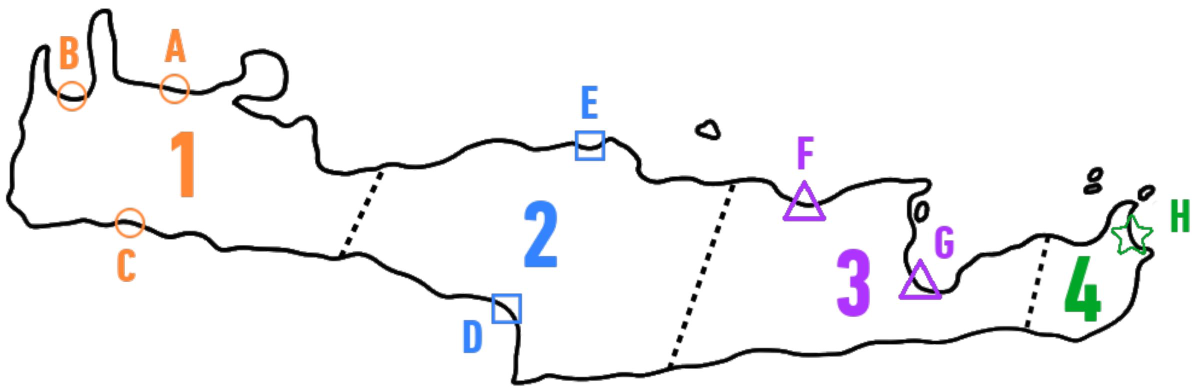


Figure 38: The Ports of Christmas Island

So what were the requirements that led to the unary and binary constraint definitions? Here they are:

Port Sizes

Not all ports are big enough to handle the ships! There are 4 port size classifications:

1. **Class 0** - These ports can only fit one ship and the ship must be either the small ship (**S**) or the medium ship (**M**), the large ship (**L**) is too big. The ports that fall in Class 0 are ports **D** and **F**.
2. **Class 1** - These ports can fit any ship by itself OR the small ship (**S**) and the medium ship (**M**) together. The ports that fall in Class 1 are ports **C**, **E**, and **H**.
3. **Class 2** - These ports can fit any ship by itself OR the small ship (**S**) and the medium ship (**M**) together OR the small ship (**S**) and the large ship (**L**) together. The ports that fall in Class 2 are ports **B** and **G**.
4. **Class 3** - These ports can fit any ship by itself OR the small ship (**S**) and the medium ship (**M**) together OR the small ship (**S**) and the large ship (**L**) together OR the medium ship (**M**) and the large ship (**L**) together. The only port that falls in Class 3 is port **A**.

Region Adjacency

In order to ensure efficient routing of truck services across the island, there are some rules about how far the ships can be docked from each other. A map of the regions can be seen in Figure **??**.

1. The small ship (**S**) and the medium ship (**M**) must be assigned to ports in the **same region**. For example if the small ship (**S**) was assigned to a port in region 1, the medium ship (**M**) must also be assigned to a port in region 1.
2. The medium ship (**M**) and the large ship (**L**) must be assigned to ports in the **same region** OR **adjacent regions**. For example if the medium ship (**M**) was assigned to a port in region 2, the large ship (**L**) can be assigned to any port in regions 1, 2, and 3 but not ports in region 4.

Logistics Companies

The island is serviced by three logistics companies that handle the offloading of the ships at ports. Each port belongs to one of the three logistics companies.

1. Easy Ship - Ports **A**, **F**, and **G**.

2. Fast Transport - Ports **C**, **D**, and **E**.

3. Quick Cargo - Ports **B** and **H**.

Each of the logistics companies are providing a Christmas special group discount and since the small ship (**S**) and the large ship (**L**) belong to the same shipping company, they'd like to make use of the group discount by **docking at ports that belong to the same logistics company**. For example if the small ship (**S**) docks at a port belonging to Easy Ship, the large ship (**L**) must also dock at a port belonging to Easy Ship.