# University of Suffolk  School of Technology, Business and Arts  BSc (Hons) Computing Level: 4

Introduction to Artificial Intelligence and Data Structure Assessment 2023/2024

S267941

Contents

[Introduction 2](#_Toc169256122)

[Task 1: Search Methods 2](#_Toc169256123)

[1. Problem Description 2](#_Toc169256124)

[2. Results 3](#_Toc169256125)

[2.1 Analysis of DFS Execution 3](#_Toc169256126)

[2.2 Analysis of BFS Execution 3](#_Toc169256127)

[2.3 Analysis of A\* Execution 4](#_Toc169256128)

[2.4 Performance Comparison of Search Algorithms 4](#_Toc169256129)

[Task 2: Bayesian Networks 5](#_Toc169256130)

[1. Problem Description 5](#_Toc169256131)

[2. Result 5](#_Toc169256132)

[1. Building and Visualizing the Bayesian Network 5](#_Toc169256133)

[2. Learning the Network Structure from Data 5](#_Toc169256134)

[3. Inference 6](#_Toc169256135)

[Task 3: Machine Learning 6](#_Toc169256136)

[1. Problem Description 6](#_Toc169256137)

[2. Result 7](#_Toc169256138)

[3. Interpretation 8](#_Toc169256139)

[Conclusion 8](#_Toc169256140)

[Bibliography 8](#_Toc169256141)

# Introduction

This report details the three tasks, which involve implementing search algorithms to find the shortest path on a map, analyzing travel data using Bayesian networks, and finally predicting the sentiment of tweets using machine learning. However, the tasks involve solving complex problems using advanced techniques and comparing their performance against specific metrics. For this reason, they will be divided into different sections, and each section will describe the problems tackled, the approaches used, and the results obtained, with supporting code snippets and figures. There will also be a full analysis explaining the implementation of the search algorithms, Bayesian networks and machine learning models.

# Task 1: Search Methods

## Problem Description

Search methods are fundamental techniques in artificial intelligence (AI) and data structures, providing the backbone for problem-solving in various domains.

For this task, implementing three of the search algorithms covered in the course was required Implementing three of the search algorithms covered in the course was required for this task. After a better analysis of the Breadth-First Search (BFS), and A\* search to find the the least distance route between Ipswich and Newcastle cities on a given map that includes 11 cities and compare the performance of the algorithms in terms of the number of nodes expanded, the maximum size of the fringe and the total running time.

## Results

### Analysis of DFS Execution

#### One of the first search algorithms was DFS. It explores nodes by going as far down a branch as possible before coming back. The importance of DFS in early AI research was highlighted by Nilsson (1980), who demonstrated its applications in various problem-solving scenarios.

#### Results

After running the DFS algorithm on the city graph:

path\_dfs = dfs(city\_graph, "Ipswich", "Newcastle")

print("DFS Path:", path\_dfs)

The output is:

DFS Path: ['Ipswich', 'Norwich', 'Cambridge', 'London', 'Sheffield', ‘Manchester’, 'Leeds', 'Newcastle']

#### Interpretation

#### The route found, which goes from Ipswich to Newcastle through a series of connected cities, is valid. It visits Norwich before reaching Cambridge, then detours through Birmingham, Sheffield and Leeds before reaching Newcastle.

#### This highlights the importance of choosing the appropriate search algorithm according to the specific requirements and constraints of the problem.

### Analysis of BFS Execution

#### Unlike DFS, BFS explores all nodes at the current depth before moving on to nodes at the next depth level. According to Knuth (1975), BFS was instrumental in developing systematic ways of browsing and searching trees and graphs and formed the basis of many subsequent algorithms.

#### Results

After running the BFS algorithm on the city graph:

path\_bfs = bfs(city\_graph, "Ipswich", "Newcastle")

print("BFS Path:", path\_bfs)

The output is:

BFS Path: ['Ipswich', 'London', 'Sheffield', 'Leeds', 'Newcastle']

#### **Interpretation**

#### The BFS algorithm has successfully found the shortest route from Ipswich to Newcastle through a series of connected cities. This demonstrates its breadth traversal approach. BFS guarantees the shortest path in terms of the number of edges.

### Analysis of A\* Execution

#### Finally, among search methods, A\* has become the cornerstone of AI for its ability to handle complex path-finding problems. Hart et al (1968) showed that A\* is both complete and optimal when using an admissible heuristic.

#### Results

After running the A\* algorithm on the city graph:

path\_a\_star = a\_star(city\_graph, "Ipswich", "Newcastle", city\_coordinates)

print("A\* Path:", path\_a\_star)

The output is:

A\* Path: ['Ipswich', 'Norwich', 'Newcastle']

#### **Interpretation**

The A\* algorithm has successfully found the shortest route from Ipswich to Newcastle through a series of connected cities. It guarantees that the route found is the most optimal in terms of total cost, combining the actual distance travelled and the heuristic estimate.

### Performance Comparison of Search Algorithms

#### Results

The performance metrics collected for DFS, BFS, and A\* are presented in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Path | Nodes Expanded | Max Fringe Size | Running Time (s) |
| DFS | ['Ipswich', 'Norwich', 'Cambridge', 'London', 'Sheffield', ‘Manchester’, 'Leeds', 'Newcastle'] | 11 | 8 | 0.0012 |
| BFS | ['Ipswich', 'London', 'Sheffield', 'Leeds', 'Newcastle'] | 10 | 5 | 0.0013 |
| A\* | ['Ipswich', 'Norwich', 'Newcastle'] | 7 | 3 | 0.0009 |

#### **Interpretation**

The table shows that A\* is the most efficient search method in terms of extended nodes and execution time. The heuristic function guides the search more efficiently towards the goal, reducing unnecessary errors.

# Task 2: Bayesian Networks

## Problem Description

A Bayesian network, sometimes called a belief network or Bayesian belief network, is a type of probabilistic graphical model. It uses a directed acyclic graph (DAG) to describe a set of variables and their conditional relationships. Judea Pearl, who developed the theoretical foundations and algorithms for AI reasoning under uncertainty, made a major contribution to the development of the concept in the 1980s (Pearl, 1988).

This task uses Bayesian networks to analyze the relationship between preferred transport modes and certain socio-economic and demographic characteristics.

## Result

### Building and Visualizing the Bayesian Network

Visualising a Bayesian network starts with loading the travel survey data, which includes variables such as age, gender, education, occupation, place of residence and preferred mode of travel. This helps to understand the relationships between the different variables in the dataset.

### 2. Learning the Network Structure from Data

#### Result

To estimate the Bayesian network structure based on the travel survey data, the Peter and Clark (PC) algorithm is used, which generates the following edges in the learned structure.

('sex, education')].

Next, the learned model structure is compared with the expert model structure by identifying matching and mismatching edges:

The result is as follows:

Matching Edges (expert-based model and learned model): {('gender', 'education')} Mismatching edges (expert-based or learned model): {('age', 'education'), ('education', 'residence'), ('residence', 'travel'), ('education', 'occupation'), ('occupation', 'travel')}.

Finally, to evaluate the learned structure, the F1 score is calculated using true positives (TP), false positives (FP) and false negatives (FN).

#### Interpretation

The expert-based model includes dependencies that make sense from a domain knowledge perspective, such as the relationship between education and occupation, or residence and travel mode. However, the PC algorithm only identified the 'gender' -> 'education' relationship from the data, suggesting that this may be the most statistically significant relationship in this dataset.

### Inference

#### Result

The learned CPTs are used to answer some interesting questions about travel preferences. Here are the questions and the results:

Query: Most likely travel mode for profile {'Age': '30-40', 'Education': 'Bachelor', 'Gender': 'Female', 'Location': 'Urban'}

{‘Travel': 'Car'}

Run the same queries using the trained Bayesian network:

Query: Most likely mode of travel for profile {'Age': '30-40', 'Education': 'Bachelor', 'Gender': 'Female', 'Location': 'Urban'}.

{‘Travel': 'Car'}

#### Interpretation

To determine which Bayesian network, expert-based or estimated, gives the most accurate predictions of the preferred mode of travel, the mode of travel is predicted and calculated using both networks. The results show that both networks often predict 'car' as the most likely mode of travel.

However, the accuracy scores indicate that the expert-based Bayesian network has slightly higher accuracy (0.8347) than the learned Bayesian network (0.8251). This suggests that the expert-based model, which incorporates domain knowledge, is slightly better at predicting the preferred mode of travel than the purely data-based model.

# Task 3: Machine Learning

## Problem Description

Machine learning has its roots in the early days of computer science and AI research. The concept began to take shape in the 1950s, when pioneers such as Alan Turing proposed the idea of machines capable of learning from experience (Turing, 1950). In 1959, Arthur Samuel defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959).

For this task, a machine learning model is built to predict the sentiment of tweets with a given dataset divided into a training set of 200,000 tweets and a test set of 20,000 tweets.

## Result

Data Preprocessing

Data Shapes:

* Training set: 160,000 samples, 3,000 features
* Validation set: 40,000 samples, 3,000 features
* Test set: 20,000 samples, 3,000 features

Building a Naive Bayes Model

**Validation Set Metrics:**

* Accuracy: 0.7453
* Precision: 0.7562
* Recall: 0.7309
* F1 Score: 0.7433

Evaluating the Naive Bayes Model

**Test Set Metrics:**

* Accuracy: 0.7498
* Precision: 0.7543
* Recall: 0.7410
* F1 Score: 0.7476

Running Another Machine Learning Algorithm (Logistic Regression)

**Validation Set Metrics (Logistic Regression):**

* Accuracy: 0.7528
* Precision: 0.7398
* Recall: 0.7865
* F1 Score: 0.7624

**Test Set Metrics (Logistic Regression):**

* Accuracy: 0.7600
* Precision: 0.7388
* Recall: 0.8043
* F1 Score: 0.7701

## Interpretation

The comparison between the two models shows that logistic regression has a slight advantage over naive Bayes in this task. Although both models perform well, the higher recall of logistic regression indicates that it is more effective at identifying positive sentiments.

In conclusion, logistic regression is recommended for this task due to its better performance measures.

# Conclusion

This report presents a comprehensive analysis of three different tasks involving search algorithms, Bayesian networks and machine learning models. By implementing and comparing different approaches, the study has highlighted the strengths and limitations of each method. The search methods demonstrated the effectiveness of A\* search, the Bayesian network analysis provided insight into travel behavior, and the machine learning task demonstrated the superiority of logistic regression over naive Bayes in sentiment analysis. Ultimately, these results highlight the importance of selecting appropriate techniques and tools to solve complex problems in AI and data science.

# Bibliography

Guo, H., Tan, C.H., Cheung, Y.M. and Liu, J. (2007). Bayesian network models for data fusion. Information Fusion, 8(2), pp.117-128. [Accessed: 28 Mai. 2024]

Jordan, M.I. (1998). Learning in graphical models. NATO Science Series D: Behavioural and Social Sciences, 89, pp.1-31. [Accessed: 28 Mai. 2024]

Koller, D. and Friedman, N. (2009). Probabilistic graphical models: Principles and Techniques. Cambridge, MA: MIT Press. [Accessed: 28 Mai. 2024]

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. Proceedings of the 7th Conference of the Society for Cognitive Science, pp. 329-334. [Accessed: 28 Mai. 2024]

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference. San Mateo, CA: Morgan Kaufmann. [Accessed: 28 Mai. 2024]

Russell, S.J. and Norvig, P. (2021). Artificial Intelligence: A Modern Approach. 4th ed. Hoboken, NJ: Pearson. [Accessed: 28 Mai. 2024]