National College of Ireland

National College of Ireland

MSc in Data Analytics

Release time: 09:00 January 4th 2025
Submission time: 21:00 January 4th 2025

Duration: 12 hours

Terminal Assessment-Based Assignment
Statistics & Optimisation

Instructions

1. You are allowed to use your class notes. If the lecture notes are used as a resource for this assignment, do not copy them directly but rather paraphrase them (write in your own words).
2. **You are not allowed to discuss your solution with other students during the examination. If it is found that a student has discussed his/ her solution with other students, the case will be referred to disciplinary committee for further actions.**
3. This is a Turnitin assignment, and the plagiarism will be checked based on Turnitin database. It will be used to check whether a text is copied from Internet, any other source or peer students.
4. You should need to submit the solution of this assignment on your Moodle page at the TERMINAL ASSIGNMENT Link. You should submit the PDF (or DOC) file by the end of your exam time.
5. You can also draw a diagram/ figure on a piece of paper, take a picture with your mobile phone camera, and submit the photo on Moodle. Please ensure that you write below your drawing the question number the drawing corresponds to.
6. Use a single-column layout document.
7. Font size for the body of the text should be 12-point Times New Roman/ Arial.
8. Include student name, student ID and course name at the top of the first page.
9. The question number being addressed must be clearly indicated in the document.
10. You should submit your code separately in the link provided.

Attempt **all** 3 Questions

Part 1:

A delivery person needs to visit a set of cities to deliver packages. Each city has a **time-based restriction**, a **value** for the package delivered, and optional **item-based conditions**. The goal is to determine the optimal route that maximizes the total value of delivered packages while satisfying all constraints. The first step is to generate the distances as in Table 1. You need to use the attached file where you set your seed value to the last four digits of your student ID. The speed of the travel is 30 km/hour so the time it takes to travel between two cities will be the distance divided by the speed. The delivery person must start at city A while each city can be visited in the following time windows shown in Table 2.

Table1: Distances between cities

From/to	A	B	C	D	E	F	G	H
A	0	2	5	8	6	4	7	9
B	2	0	3	4	7	5	6	8
C	5	3	0	2	3	6	5	7
D	8	4	2	0	6	8	7	9
E	6	7	3	6	0	5	4	6
F	4	5	6	8	5	0	2	3
G	7	6	5	7	4	2	0	4
H	9	8	7	9	6	3	4	0

Table 2: Time constraint

City	Time Window	Delivery value
B	9:00-12:00	500
C	10:00-14:00	300
D	11:00-13:00	400
E	13:00-17:00	600
F	8:30-10:30	200
G	12:00-16:00	350
H	14:00-18:00	450

In addition to the time constraint there are the following constraints:

- You can either visit city D or C or not both.
- You can either visit city H or G or not both.
- Due to fuel constraints, you can visit a maximum of five cities.
- The delivery person should visit at least three cities
- If the delivery person visit city B then they must visit F.

a) Define the decision variables and formulate the mathematical model for the above problem

(10 marks)

b) Solve the problem using the Exact methods (Pulp) or using a heuristic method

(13 marks)

c) Remodel the problem so that we have two objective functions one to maximise the profit and the other to minimize the total distance travelled

(7 marks)

d) Describe the different steps followed in NSGA II algorithm and use it to Solve the Multi-objective optimisation problem formulated in the previous question.

(10 marks)

Part 2: Logistic Regression

You are tasked with building a logistic regression model to predict whether a tumor is malignant or benign using the Breast Cancer Dataset. The dataset (cancerdata.csv)includes 30 numeric features and a binary target variable (`0` for benign, `1` for malignant).

1. Write the logistic regression equation relating independent variables to dependent variable y.
   (3 marks)
2. Load the Breast Cancer Dataset and perform an exploration analysis of the data. Is there any class imbalance?
   (3 marks)
3. Use both independent variables and software to compute the estimated model equation by:
   I. Split the data into training and testing.
   II. Show the evaluation metrics of your model
   III. Present interpretation of your model
   (10 marks)
4. Use $\alpha = 0.05$ to determine whether each of the independent variables is significant.
   (4 marks)
5. What is the estimated odds ratio for five variables of your choice? Interpret it.
   (5 marks)
6. A patient's tumor has the following feature values (use mean radius, mean texture, mean smoothness, and mean symmetry):

   Mean radius = 14.5

   Mean texture = 18.0

   Mean smoothness = 0.095

   Mean symmetry = 0.180

Use the trained model to predict whether this tumour is likely to be malignant or benign. Show your work and explain the result. **(5 marks)**

Part 3: Time series

The file times-series.csv, available on Moodle, is a quarterly time series of historical sales data for different products. The variable you should use is determined by the last digit of your student number, as shown below:

Last digit	Variable to use
0-1	V2
2-3	V3
4	V4
5	V5
6	V6
7	V7
8	V8
9	V9

1. Perform a preliminary assessment of the nature and components of the raw time series, using visualisations as appropriate.

   (10 marks)

2. Estimate suitable time series models from the categories covered in class and assess the adequacy of each model.

   (15 marks)

3. Using the best model in the previous question, perform forecasting for the next four quarters.

   (5 marks)