AUT

COMP809 – Data Mining & Machine Learning

Semester 1, 2024

Assignment 1

Maximum Marks: 100 22 March 2024

Paper Description:	Data Mining & Machine Learning
Paper Code:	COMP809
Total Marks:	100
Date:	22 March 2024
Deadline:	12 April 2024

INSTRUCTIONS:

1. This is an individual assignment.
2. Only documents in pdf or html will be accepted. You can generate the document directly on Jupyter Notebook.
3. Submit the pdf or html file via Canvas.
4. Formats other than pdf will be ignored and the author will be asked to re-submit the assignment. Resubmissions will be subject to the late policy outlined in the study guide (i.e., 5% per day up to 5 days).
5. The python code required to complete this assignment, which includes code to support your conclusions & answers, must be embedded in the document in the corresponding answer as text (not image), unless otherwise specified. This code will be marked. Unsolicited python scripts submitted separately will not be marked.
6. Read carefully and answer all the questions as requested. Any material or information unrelated to the correct answer may result in a significant reduction of marks for that question.
7. Do not forget to fill in and sign the cover sheet which must be the very first page in the pdf. Use software such as Adobe Acrobat Pro on the Uni computers to include the file at the start of your document.
8. If you need an extension or if your performance has been impacted by some extenuating circumstances, then you must complete a special consideration form on Canvas.
9. Only the techniques studied in this course will be accepted.
10. The comprehension of the questions is part of the assignment.

Grade table:

Question:	1	2	3	Total
Points:	30	35	35	100
Score:				

QUESTIONS:

1. Nitrogen Oxides are a family of poisonous, highly reactive gases. These gases form when fuel is burned at high temperatures. NOx pollution is emitted by automobiles, trucks and various non-road vehicles (e.g., construction equipment, boats, etc.) as well as industrial sources such as power plants, industrial boilers, cement kilns, and turbines. NOx often appears as a brownish gas. It is a strong oxidizing agent and plays a major role in the atmospheric reactions with volatile organic compounds (VOC) that produce ozone (smog) on hot summer days. Source: United States Environmental Protection Agency−EPA.

   The `NOxEmissions.csv` file contains 8088 observations on the following 4 variables:

   - *julday*: day number, a factor with levels 373 ... 730, typically with 24 hourly measurements.
   - *LNOx*: log of hourly mean of NOx concentration in ambient air [ppb] next to a highly frequented motorway.
   - *LNOxEm*: log of hourly sum of NOx emission of cars on this motorway in arbitrary units.
   - *sqrtWS*: Square root of wind speed [m/s].

   In this study, we are interested in modelling the *LNOx* concentration as a function of the other variables. The time dependency will be omitted. The same with the space dependency if there is any.

   Total for Question 1: 30

   (a) Describe the data preprocessing. Justify your answers. [3]

   (b) Describe the distribution of the variable *LNOx*. [4]

   (c) Fit a linear model to explain the variable *LNOx* as a function of *LNOxEm* and *sqrtWS*. Comment on the model. Justify your answer. [15]

   (d) Discuss the relationship between the dependent and independent variables. Interpret in a way that someone who is not familiar with the field can understand the parameter associated to the predictor *LNOxEm*. [4]

   (e) Predict the Nitrogen Oxides concentration for a $LNOxEm = 7.5$ and $sqrtWS = 1.3$. Interpret your results in a way that someone who is not familiar with linear models can understand. [4]

2. The `nassCDS.csv` file contains data on Airbag and other influences on accident fatalities. The data has been collected in US, for 1997-2002, from police-reported car crashes in which there is a harmful event (people or property), and from which at least one vehicle was towed. Data are restricted to front-seat occupants, include only a subset of the variables recorded, and are restricted in other ways also. The data has 26,217 observations on the following 15 variables:

   - *dvcat*: ordered factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+.
   - *weight*: Observation weights, albeit of uncertain accuracy, designed to account for varying sampling probabilities.
   - *dead*: factor with levels `alive dead`.
   - *airbag*: a factor with levels `none airbag`.
   - *seatbelt*: a factor with levels `none belted`.

- *frontal*: a numeric vector; 0 = non-frontal, 1=frontal impact.
- *sex*: a factor with levels `f m`.
- *ageOFocc*: age of occupant in years.
- *yearacc*: year of accident.
- *yearVeh*: Year of model of vehicle; a numeric vector.
- *abcat*: Did one or more (driver or passenger) airbag(s) deploy? This factor has levels `deploy nodeploy unavail`.
- *occRole*:a factor with levels `driver pass`.
- *deploy*: a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed.
- *injSeverity*: a numeric vector; 0:none, 1:possible injury, 2:no incapacity, 3:incapacity, 4:killed; 5:unknown, 6:prior death.
- *caseid*: character, created by pasting together the populations sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

In this project we are interested in explaining the probability of surviving (variable *dead*) given *airbag*, *seatbelt*, *frontal*, *sex*, *ageOFocc*, *yearVeh*, and *deploy*. For this fit a proper generalised linear regression model.

Total for Question 2: 35

(a) Describe the data preprocessing. Justify your answers. [3]

(b) Is the use of the seat belt independent of whether the passenger survives or not? Justify your answer. Use only the variables related to this question in your analysis. [6]

(c) Is there a mean age difference between the following injury severity (*injSeverity*) groups: none, possible injury, no incapacity, incapacity, and killed? Justify your answer. Use only the variables related to this question in your analysis. [6]

(d) Fit a model that explains the dependent variable as a function of *airbag*, *seatbelt*, *frontal*, *sex*, *ageOFocc*, *yearVeh*, and *deploy*. Use 70% to train the model and 30% to test it. Comment on the performance of the model in a way that someone who is not familiar with the concepts can understand. [12]

(e) Interpret the parameter associated to *seatbelt* and *ageOFocc* in a way that anybody can understand. [4]

(f) Predict the odds of not surviving for the following two scenarios:

1. There is no airbag, the passenger is not wearing seatbelt, it is a frontal impact, the passenger is female 70 years old, and the airbag is not deployed.
2. There is airbag, the passenger is wearing seatbelt, it is a frontal impact, the passenger is female 70 years old, and the airbag is deployed.

Interpret your results. [4]

3. The dataset is being used in a project that aims to study the international student flow and its relationship with the level of globalisation of countries and their national higher educational systems in 2019. The data were extracted and merged from three sources: UNESCO, Swiss KOF Economic Institute Dataset, QS World University Rankings. The `data_q3.xlsx` file contains many variables, but in this study, we are interested in the following ones:

- *InboundRatio*: Inbound mobility rate for international students.
- *InternationalStudentsNO*: Total number of inbound international students.
- *KOFPoGI*: Political Globalization index.
- *KOFEcGI*: Economic Globalisation index.
- *KOFSoGI*: Social Globalization index.
- *ISCED5 Percentage*: Participation in short cycle tertiary education.
- *ISCED6 Percentage*: Participation in Bachelor's level.
- *ISCED7 Percentage*: Participation in Master's level.
- *ISCED8 Percentage*: Participation in Doctoral level.
- *top_50_count*: Number of universities ranked in the top 50.
- *top_100_count*: Number of universities ranked in the top 100.
- *top_500_count*: Number of universities ranked in the top 500.
- *top_1000_count*: Number of universities ranked in the top 1000.
- *WESP*: World Economic Situation and Prospects 2021.
- *country_x*: country.

Notes: ISCEDs variables refer to the participation percentage by level of higher education. It is calculated by the number of students enrolled in the corresponding level divided by the population of the official age for tertiary education, multiplied by 100. About *KOFPoGI*, *KOFEcGI*, and *KOFSoGI* indexes, the higher the better. The inbound mobility rate is the number of students from abroad studying in a given country, expressed as a percentage of total tertiary enrolment in that country.

Total for Question 3: 35

(a) Describe the data preprocessing. Justify your answers. [3]

(b) Perform an exhaustive K-mean cluster analysis on the variables of interest. How many clusters do you propose? Justify your answer. [15]

(c) Perform an agglomerative cluster analyses. How many clusters do you propose? Justify your answer. [12]

(d) What do you conclude? Provide an interesting remark(s). Justify your answers. (3-4 sentences). [5]
