UNIVERSITY OF LONDON

BSc Computer Science

Module: CM3005 - Data Science

Introduction

This coursework requires you to utilise a linear regression algorithm and to apply techniques from data visualisation tools and statistical analysis. You will need to identify a suitable domain-specific problem area and an associated data set. This implies that the selected domain is expected to have data with a linear trendline. Hint. You will need to visualise data as your first step before deciding the domain/dataset. Also, it is advisable to select a dataset which has less than 10,000 entries so you can load and process it faster.

When you have your dataset selected you will then proceed with the preprocessing of data where you check for imperfections and perform normalisation. Remember that your dataset should have missing data, come in several csv files etc., to make use of the preprocessing step. Do not choose a dataset which is in 1NF.

When your dataset is ready you can proceed with the implementation process. That involves building and testing the model and obtaining and evaluating the results. We will award some extra points for those who choose to implement this assignment further. Discuss any additional work clearly in your Jupyter notebook i.e. *Additional work* section or similar. The project is expected to be developed by using the Python programming language and Jupyter notebook. Provide well-commented Python code accompanied by documentation describing the following steps:

Task/steps

1. Domain-specific area and objectives of the project (200-500 words)

The first step of the coursework is to identify and describe the domain-specific area. This is an area of industry or science where a linear regression model will contribute. It can be any field which can benefit from machine learning. Then state and justify the objectives of the project. Discuss its impact and contribution to the domain-specific area. State any contribution which the results may make to the challenge addressed.

Examples: Housing prices in the UK (United Kingdom). To show how prices were affected by Brexit in London and Manchester. Life expectancy globally. To show the effect of GDP per capita.

2. Dataset description (200-500 words)

The next step is to identify a suitable dataset representative for the coursework's domain. The use of a suitable dataset can address all the steps outlined in this assignment. Provide a description of the dataset, its size, data types, the way the data was acquired. State clearly the source of the dataset. Kaggle.com provides a wide variety of datasets.

Example: Housing market dataset, collected by the US Census Bureau providing information about real estate in California. The dataset can be accessed via the Kaggle official website.

3. Data preparation (acquisition/cleaning/sanitisation/normalisation)

Convert/store the dataset locally and preprocess the data. This is usually equivalent to transforming a table from a database into First Normal Form (1NF). Describe the preprocessing steps and why they were needed. Describe the file type/format, for example CSV file. Process the dataset for missing data if needed. For this part use Pandas DataFrame.

4. Statistical analysis

Identify key series of the dataset and provide statistical summary of the data, including:

- Measures of central tendency
- Measures of spread
- Type of distribution

This can be done by using libraries such as NumPy, pandas and SciPy. Most likely the dataset will consist of multiple series.

5. Visualisation

Visualise key data series within the dataset by using the appropriate graphs. This can be done by using Python libraries, such as Matplotlib. Accompany any diagram with explanations - we will not award any points if you do not. Draw conclusions based on the diagrams, which otherwise, without visualisation would be difficult or impossible. Which visualisation (of the ones you prepared) is most important and why?

6. Build your ML (Machine Learning) model

- Identify the features and the labels which will be used in the data regression model and justify why they were selected
- Explain their (i.e. features) importance for the process of building the ML model
- Build the model by using an appropriate Python library, such as Weka or Scikit-learn

- Run and evaluate your model: does your model fit the data or further pre-processing is required?

7. Validation

Consider how the model's results could be validated using either cross validation or other models/ensembles.

8. Feature engineering

Your model would benefit from using feature engineering techniques or polynomial features as described in Topic 5.4. Implement these techniques and re-evaluate your model.

9. Programming style

The Python code is expected to meet certain standards as described by most coding conventions. This includes code indentation, not using unnamed numerical constants, assigning meaningful names to variables and subroutines. Additionally, the code is expected to be commented, including all variables, sub-routines and calls to library methods.

10. Evaluation of your model (200 to 400 words)

Evaluate the results of the machine learning model:

- Numerically evaluate the performance of the model
- Justify the use of an appropriate measure such as RMSE
- Provide a reflective evaluation of the developed project considering the obtained results
- Describe its contributions to the selected domain-specific area
- Discuss whether the solution is transferable to other domain-specific areas

Deliverables

1. Your Jupyter notebook including the sections of the report starting with questions 1 and 2. Include your code and results as produced from the Jupyter notebook and conclude with question 10 to evaluate your work. This should be saved as a single .**pdf** file and then uploaded to the first prompt.
2. Your code (i.e. your Jupyter notebook code as in 1.) in .**txt** format. Use Download/Export menus on Jupyter notebook, Jupyter Lab, Colab etc. Then upload the code in the second prompt. Please see submission page for further details. **This is a submission requirement as we need this to run your code. You will receive zero points for the entire project if you fail to submit your code in this way.**

Rubric

Marks are shown in square parentheses.

Q1. Describe the domain-specific area and the objectives of the project. Why can we use a linear regression model in this domain and what are you trying to achieve? (200-500 words)
[0] Missing or incorrect
[1] Briefly discussed i.e. couple of lines
Then +1 for each [overall 5 points for Q1]

Domain-specific area clearly stated with informative description
Objectives clearly stated with sufficient details
Potential contributions included
Fully referenced work included

Q2. Description of the selected dataset. Why this dataset fits the above purpose? (200-400 words)
[0] Missing or incorrect
[1] Briefly described

Then +1 for each [overall 6 points for Q2]

- Size
- Origin
- Structure and data types

Fitness for LR [2 points]

Q3. Pre-processing
[0] Missing or incorrect
[1] Briefly described and no use of pandas
[2] Working code fragments with some pre-processing steps, dataset not in 1NF
[3] All necessary pre-processing steps undertaken, dataset in 1NF, with fixes to missing values, cleaning data achieved

Q4. Statistical Analysis
[0] Missing or incorrect
[1] Briefly described
[2] Adequately described, some statistical information provided, the mean, median and standard deviation
[3] Dataset accurately summarized by providing the appropriate statistical data. That includes the above plus skewness and kurtosis
[4] Analysis of findings e.g., which measure is more interesting and why?

Q5. Data Visualisation
[0] Missing/incorrect/only visuals included without any discussion
[1] Basic description, some visualisation without conveying the properties of the data
Then +1 for each [overall 4 points for Q5]

Informative visualisation of key variables using appropriate types of graphs.
Visualisations alongside informative text
Student discusses the most important visualisation

Q6. Building ML model
[0] Missing or incorrect
[1] Briefly described
[2] Working solution with unconfirmed results
[3] Working solution with confirmed results generated and presented appropriately

Q7. Validation
[0] None attempted

[1] Briefly described but not attempted in code
[2] Cross validation used but not reflected on results
[3] Cross validation with confirmed results generated and presented appropriately

Q8. Feature engineering
[0] None attempted
[1] Briefly described
[2] Working solution with unconfirmed results
[3] Working solution with confirmed results generated and presented appropriately

Q9. Programming style
[0] No code provided
[1] A minimal attempt at readability
[2] The source code is readable with some comments
[3] The source code is of high quality and follows general coding convention
[4] Professional level code presentation (e.g., extensive and consistent approach to documentation through comments, consistency of approach - camelCase all the way through)

Q10. Evaluation of the project and its results (200 to 400 words)
[0] Missing or incorrect
[1] Briefly described
[2] Some conclusions without numerically evaluating the ML model and the results
[3] Results discussed with the performance of the model numerically evaluated
[4] Numerically evaluated and the measure is justified
[5] Detailed project evaluation (functionality, results, transferability, contributions)

Q11. Overall bonus points for ambition and originality
10 points overall (5 points for ambition and 5 for originality)