Coursework Assignment: Building a Data Regression Model

University of London
BSc in Computer Science
Data Science

Please note: This is a graded assignment. This is a draft of the assignment made available before course work submission is open, only minor changes may occur.

Contents

I. Introduction

This coursework requires you to develop a machine learning model 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.

1. Domain-specific area

The first step of the coursework is to identify and describe the domain-specific area. This is an area of industry or science, where the regression model will contribute. It can be any field, which can benefit from a machine learning model.

2. Dataset

The next step is to identify a suitable dataset which is representative for the domain of the coursework. A suitable dataset is able to addresses all steps outlined in this assignment and would allow maximum marks. Provide a description of the dataset, its size, data types, the way the data were acquired. State clearly the source of the dataset. Large technology companies, such as Microsoft, Google and Amazon, provide 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. Objectives of the project

State and justify the objectives of the project, that is the machine learning model accompanied with some visualisation tools and statistical analysis. Discuss its impact and contributions to the domain-specific area. State any contributions, which this project can bring to the domain-specific area.

II. Implementation

This part of the coursework is the implementation of the project. It includes storing and preprocessing of the dataset, building and testing the model and obtaining and evaluating the results. The project is expected to be developed by using the Python programming language and Jupyter notebook. Provide well-commented Python code accompanied by document describing the following steps:

4. 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.

5. Most likely the dataset will consist of multiple series. 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.

6. 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. Draw conclusions based on the diagrams, which otherwise, without visualisation, would be difficult or impossible.
7. Identify the features and the labels, which will be used in the data regression model and justify why they were selected. Explain their importance for the process of building the ML model. Build a regression model by using appropriate Python library, such as Weka or Scikit-learns. Describe the employed ML algorithm, for example Random Forest or Support Vector Machine, and the reasons for choosing it. Run and test the machine learning model.

8. 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, that is every variable, sub-routine and a call to a library method to be described.

III Conclusions

9. Evaluate the results of the machine learning model. Use measures, such as RMSE, to numerically evaluate the performance of the model.

10. Provide a reflective evaluation of the developed project in light of the obtained results. Describe its contributions to the selected domain-specific area. Discuss whether the solution is transferable to other domain-specific areas. Discuss whether the project can be reproduced by using different programming languages, development environments, ML libraries and ML algorithms. Review the possible benefits or drawbacks of choosing different approaches.

Rubric

Marks are shown in square parentheses.

I. Introduction

1. Introduction to the domain-specific area (200-500 words)
[0] Missing
[5] Briefly discussed
[10] Domain-specific area clearly stated, informative description, referenced work

2. Description of the selected dataset (200-500 words)
[0] Missing
[5] Briefly described
[10] Described in sufficient details, including origin, size, structure, data types

3. Objectives of the project (200-500 words)
[0] Missing
[5] Briefly discussed
[10] Objectives clearly stated with sufficient details and possible contributions

II. Coding (commented source code)

4. Pre-processing
[0] Missing
[5] Working code fragment with some pre-processing steps, dataset not in 1NF
[10] All necessary pre-processing steps undertaken, dataset in 1NF

5. Statistical Analysis
[0] Missing
[5] Some statistical information provided
[10] Dataset accurately summarised by providing the appropriate statistical data

6. Data Visualisation
[0] Missing
[5] Some visualisation without conveying the properties of the data
[10] Informative visualisation of key variables using appropriate types of graphs

7. Building ML model
[0] Missing
[5] Working solution with unconfirmed results
[10] Working solution with confirmed results generated and presented

8. Programming style
[0] Unmeaningful names in code, no comments, use of 'magic numbers'
[5] The source code is readable with some comments
[10] The source code is of high quality and follows general coding convention

III. Conclusions

9. Results of the ML model (100 to 300 words)
[0] Missing
[5] Results discussed but not numerically evaluated
[10] Results discussed with the performance of the model numerically evaluated

10. Evaluation of the project and its results (200 to 400 words)
[0] Missing
[5] Some conclusions without evaluating the ML model and the results
[10] Detailed project evaluation (ML model, functionality, results, reproducibility)