CM3015 Mid-term coursework

This coursework is in the form of a mini-project. You are asked to apply at least two machine learning algorithms to a dataset or datasets of your choosing. You will write your findings in a report. The report, written as a document and in a formal style, and your Jupiter notebook, constitute the hand-in.

Aim

You must decide what you wish to achieve. Possibilities include: the best ML model for a particular dataset; comparison of algorithms across differing datasets; systematic examination of variations of a particular algorithm (for example, naive k-means vs PCA initialisation); comparison of different algorithms on a dataset or datasets (for example, k-means vs other clustering algorithms).

Choosing a dataset/datasets

You can either choose a dataset/datasets that is/are packaged with a machine learning library or pick a dataset that interests you from a public repository such as kaggle.com.

For example, scikit-learn contains several standard, classic datasets such as Iris, Wine and Hand-written digits. These are perfect for this project.

You might wish to browse (e.g.) kaggle for an interesting dataset but please ensure that you can vectorise the dataset into a suitable form for input into a machine learning algorithm. You will not receive any credit for manipulating the data prior to analysis.

Algorithms

You should apply at least two machine learning algorithms from the first part of this module to your chosen problem. Specifically, at least two from: kNN, decision trees, linear regression, gradient descent, polynomial regression, Bayesian classification, k-means and PCA.

You should implement at least one of the ML algorithms from scratch. This/these implementation(s) must be in standard Python code and should not refer to any machine learning libraries. The use of numpy and matplotlib is permissible (and expected).

Methodology, Analysis and Evaluation

The first half of this module (Topics 1-6) introduced several important ML techniques such as Training/test set splitting, classifier evaluation metrics (precision, accuracy, …), data scaling, over/under fitting, regularisation and cross-validation. You should utilise these techniques wherever appropriate.

The report

The report must be structured as follows.

1. Abstract. This is a single paragraph that summarises the aim and findings of your project. [5 marks]

2. Introduction. This section places your project in a machine learning context as exemplified by the academic literature. You can amplify the aim as stated in the abstract and explain why this aim is interesting and relevant. You might introduce your dataset(s) and explain any particular problems both with the dataset and with known investigations of this dataset. [10 marks]

3. Background. Here you should explain in *your own words* how your algorithms work. This is where you can demonstrate that you understand the tools and models that you used. Marks will be gained accordingly. [20 marks]

4. Methodology. Set out how you explored the dataset and/or algorithm modifications. For example, you might have decided to use cross-validation; if so, explain why this technique was necessary. [20 marks]

5. Results. Your results must be stated clearly. Tables are recommended. You should cross-reference to the experiments that you described in the Methodology section. [10 marks]

6. Evaluation. This is a chance to demonstrate a critical awareness of the strengths and weaknesses of your project. Remember that a research project is judged by referring to the stated aim. A project does not have to succeed! Marks are allocated for how you undertook the project and for understanding and insight. A very ambitious aim might be unachievable within the terms of this project. [20 marks]

7. Conclusions. State succinctly your findings and how they relate to your aim. [10 marks]

8. References. List any academic work (such as a book or a research paper) that you refer to in the main body of the report. [5 marks]

Jupyter notebooks

All your work must be in a Jupyter notebook. Submit this notebook as an html file. *Do not submit your .ipynb file.* We will not run your source code but we will definitely wish to read it!