Coursework (CM3111)
Big Data Analytics

Eyad Elyan, *e.elyan@rgu.ac.uk*

October 11, 2018

Overview

The purpose of this coursework is to help you understand, identify, apply and evaluate machine learning algorithms. You will work with data, design, implement and evaluate solutions to mine data and extract knowledge. You will also use apply different visualisation methods to communicate results.

You will will use a machine learning algorithm/s to build a model for classifying a dataset of your own choice, as outlined in the tasks below. Other learning algorithms are possible (i.e. unsupervised machine learning). Please read the instructions below very carefully, so that you understand what is required.

1 Data Exploration

In this task, you will chose a dataset, define the main objective of your experiment and review (briefly) relevant literature. You will also apply basic data exploration techniques, produce some visuals/ diagrams that explain the data. These steps are outlined below:

1.1 Dataset Choice

First, you will select a dataset related to a real world problem that best suits your area of interest. There are abundant of websites that provide publicly available datasets. A categorised list of datasets from GitHub can be found at https://github.com/caesar0301/awesome-public-datasets. The UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets.html is another long standing source of benchmark datasets for data mining and machine learning research. Kaggle https://www.kaggle.com/datasets has interesting real world problems and datasets. In this task you need:

- Clearly state the name and source of the data
- Explain briefly why you chose this dataset

1.2 Problem Statement & Data Exploration

Once you have chosen your set, you need to provide a brief description to the data set and write script for quick exploratory analysis to the data. In particular:

1. Brief description of the data set (i.e. the dataset contains records for patients with heart disease.... The aim is to build a predictive model to predict if a certain person has a heart disease or not, ...)
2. Clearly define the main objective of your work (i.e. This dataset will be used to build a model to predict if an email is spam ore not?)

3. Names of the features (columns)
4. Class/ label distribution in the dataset

Your code must be commented and the steps explained. Visuals and diagrams provide an excellent way for describing data, you need to at least produce two diagrams that explains the dataset you have chosen (i.e. Class distribution, features correlations, etc...).

1.3 Pre-processing

Most data related problems require some form of data pre-processing. In this task, and upon completing exploring your data, you need to decide what pre-processing steps need to be applied for your dataset, and do so. In particular:

- If you have some missing values in your data, you may apply some missing-values handling methods. You need to show how did you check for any missing values and how you handle it
- You may will need to standarise/ normalise your data to ensure better results
- Generate new features, drop irrelevant features, etc...
- Any other pre-processing steps that you think is necessary to apply for your chosen dataset (i.e. dimension reduction, etc...)

Again, your code must be commented and the steps explained. This is very important when applying some data pre-processing methods so that results can be reproduced.

2 Modelling/ Classification

In this task you will build a classification model to classify/ predict the class label in your dataset. The model choice is up to you. In other words you can use SVM, Random Forest, Neural Net or any other classification model that you studied in the lectures. You must justify your choice of the method, and write, explain and comment the code you produce. More specifically, your code must:

1. Divide the set into training and testing subsets
2. Build a model of your choice using the training set
3. Test and evaluate your model
4. Report and discuss results

You must discuss training and testing error, and any other issues related to results. Code must be commented and the steps explained. Your code must be developed as a set of functions (i.e. a function to divide the dataset into train/test, function to build a model, function to return results, ...).

3 Improving Performance

In this task you will improve the performance of your model in the previous task. There are different ways to improve and fine tune your model. Please refer to the lecture notes. More specifically, you may want to explore one or more of these options

1. Fine tune the parameters of your model, or search for some existing techniques to optimise these parameters

2. Use different metrics for evaluating the model (i.e. perhaps the data is not balanced, and you might want to trying some other methods)
3. Try different models and compare the results
4. Change the partitioning of the dataset. In some cases more training data may help
5. Use cross-validation if not used already,
6. Any other methods you think might be appropriate to use

You must discuss, explain and justify your approach and attempt to improve model performance. In other words you need to demonstrate that you have made an informed choices while attempting to improve the model performance.

4 Reproducing Results

Your solutions to this course work needs to be produced as a *.Rnw* and compiled into a PDF file. In other words, you will end up with one PDF file that includes your explanation, coding and output results (diagrams, tables, etc...). Please refer to week 1 materials for more information about reproducing results. A template will be provided for you as well.

Deliverables & Due-Dates

By the **Friday, 30^{th} of November, 2018**, *5pm*, you must submit your solutions for all parts of the coursework. Your submission must be in the form of a PDF file (a compiled version of your *.Rnw* file. Following the steps and code provided in your PDF file, I should be able to reproduce your results. Make sure you comment your code, and explain your approach where possible.

Note: Grading Scheme is in the following pages.

Grading Scheme

- ***Data Exploration***

A Excellent choice of dataset and very clear set objectives are outlined. Data exploration and pre-processing is thorough, in-depth, and covers all expected areas. Very clear justification of the chosen technology/s and excellent use of visuals.

B Very good choice of dataset and clear set objectives are outlined. Data exploration and pre-processing is almost thorough, in-depth, and covers almost all expected areas. Very good justification of the chosen technology/s and very good use of visuals.

C Good choice of dataset and clear set objectives are outlined. Data exploration and pre-processing is almost thorough, in-depth, and covers most expected areas. Good justification of the chosen technology/s and good use of visuals.

D Dataset chosen for the task is satisfactory and some clear set objectives are outlined. Adequate data exploration and pre-processing, and covers some expected areas but lacks depth. Some justification of the chosen technology/s and adequate use of visuals.

E Not very adequate technical quality, lacks depth and technical understanding.

F Very minimal effort has been made. None of the expected areas were covered.

- ***Modelling/ Classification***

A Excellent work that satisfies all of the expected requirements with the expected outcomes. all of the code follows sensible layout & coding style. Methods used are well explained and justified. All functions (or sections of code) are correctly implemented, commented & and explained. Results produced are accurate and reproducible.

B Very good work that satisfies a substantial part of the expected requirements. Most of the code follows sensible layout & coding style with only minor errors. Most individual functions (or sections of code) are correctly implemented, explained and commented.

C Good attempt that satisfies the basic requirements plus some outcomes/ results are correct. Most of the code follows sensible layout & coding style with minor errors. Most individual functions (or sections of code) are correctly implemented although with some inefficiency and lack comments, explanation and justification in some parts.

D Satisfies the basic requirements. Majority of the code follows sensible layout & coding style with one or more major errors. Majority of functions (or sections of code) are correctly implemented but with some inefficiency and lack comments, explanation and justification in some parts.

E Satisfies only a very limited subset of the basic requirements. Some of the code follows sensible layout & coding style but with many errors and inaccurate outcomes. Some functions (or sections of code) are correctly implemented. Little explanation and justification have been provided

F Very minimal effort has been made and doesn't satisfy any requirement.

- ***Improving Performance***

A Excellent work that satisfies all of the expected requirements with the expected outcomes with significant improvement of results over the previous attempt. all of the code follows sensible layout & coding style. Methods used are well explained and justified. All functions (or sections of code) are correctly implemented, commented & and explained. Results produced are accurate and reproducible.

B Very good work that satisfies a substantial part of the expected requirements with improvements of results over the previous attempt. Most of the code follows sensible layout & coding style with only minor errors. Most individual functions (or sections of code) are correctly implemented, explained and commented.

C Good attempt that satisfies the basic requirements plus some outcomes/ results are correct with some improvement of results over the previous attempt. Most of the code follows sensible layout & coding style with minor errors. Most individual functions (or sections of code) are correctly implemented although with some inefficiency and lack comments, explanation and justification in some parts.

D Satisfies the basic requirements. Majority of the code follows sensible layout & coding style with one or more major errors (no improvement of results). Majority of functions (or sections of code) are correctly implemented but with some inefficiency and lack comments, explanation and justification in some parts.

E Satisfies only a very limited subset of the basic requirements. Some of the code follows sensible layout & coding style but with many errors and inaccurate outcomes. Some functions (or sections of code) are correctly implemented. Little explanation and justification have been provided

F Very minimal effort has been made and doesn't satisfy any requirement.

- ***Results Presentation***

A Excellent presentation and communication of the results using a reproducible document (i.e. *r-sweave* or *R-Markdown*) that satisfies all requirements. Proper naming of the sections, labelling of the output (i.e. figure/ table captions).

B Very good presentation and communication of the results using a reproducible document (i.e. *r-sweave* or *R-Markdown*) that satisfies substantial part of the requirements. Proper naming of the sections, labelling of the output (i.e. figure/ table captions).

C Good presentation and communication of the results using a reproducible document (i.e. *r-sweave* or *R-Markdown*) that satisfies most parts. Some proper naming of the sections, labelling of the output (i.e. figure/ table captions).

D Adequate presentation and communication of the results using a reproducible document (i.e. *r-sweave* or *R-Markdown*) that satisfies some parts.

E Inadequate and partial presentation and communication of the results using a reproducible document (i.e. *r-sweave* or *R-Markdown*). The document provides little explanation and justification of the chosen methods and techniques.

F Very minimal effort has been made and no reproducible document was provided .

The overall assignment grade is determined from the above profile based on the weighted average of the main four major parts of the coursework as follows :

- 4 Main parts in Total = Data Exploration (30%) + Modelling/Classification (30%) + Improving Performance (20%) + Results Presentation (20%) = 100% 20 = 5%

Weighted Average

- Data Exploration 30% = 3 Units
- Modelling/Classification 30% = 3 Units
- Improving Performance 20% = 2 Units
- Results Presentation 20% = 2 Units

Weights

- A = 5, B = 4, C = 3, D = 2, E = 1, F = 0

Example

- Data Exploration **A**
- Modelling/Classification **B**
- Improving Performance **A**
- Results Presentation **B**

Final Grade =

$$\frac{(\mathbf{5}\times 3)+(\mathbf{4}\times 3)+(\mathbf{2}\times 5)+(\mathbf{2}\times 4)}{10} = 4.5\ (\mathrm{A})$$