Applied Machine Learning (ICS-5110)

Coursework Specifications

Date of Issue: 19th November 2019

Deadlines: 20/01/2020 (Assignment) 24/01/2020(Homework Portfolio)

Contact person: Dylan Seychell (email: dylan.seychell@um.edu.mt)

This coursework is worth 100% of the total, final mark for this unit. You are expected to allocate approximately 75 hours to complete the assignment. You will be required to demonstrate (and be able to discuss) your working solutions in a 15-minute demo. The deadline for this assignment is Monday, 20th January 2020 at noon. The portfolio should be submitted by Friday 24th January 2020. Late submissions will not be accepted. You may find more details below. Questions regarding the assignment should only be posted in the Assignment VLE forum (and not via personal correspondence with the lecturer of this study-unit). Under no circumstances are you allowed to share the design and/or code of your implementation outside your group. The homework portfolio is an individual task and therefore you are not allowed to share any deliverable with other students. You may not copy code from internet sources, you will be heavily penalized if you do so. The Department of Artificial Intelligence takes a very serious stand on plagiarism. For more details refer to the plagiarism section of the Faculty of ICT website.

Assignment (80% of unit marks)

1. You can work in groups of 3. Equal contribution will be assumed
2. The 80 points are divided into 80% for the quality of the project and 20% for the quality of the demonstration
3. The aim of this project is to introduce you to machine learning (ML) techniques applied to real data sets. You are expected to:
    - Choose a data set from the list below:
        - http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data
        - http://archive.ics.uci.edu/ml/datasets/Air+Quality
        - http://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset
        - http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
        - http://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly
        - http://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection
        - http://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29
        - http://archive.ics.uci.edu/ml/datasets/WESAD+%28Wearable+Stress+and+Affect+Detection%29

        - http://archive.ics.uci.edu/ml/datasets/BLE+RSSI+Dataset+for+Indoor+localization+and+Navigation
        - http://archive.ics.uci.edu/ml/datasets/Container+Crane+Controller+Data+Set
        - http://archive.ics.uci.edu/ml/datasets/Parkinson+Disease+Spiral+Drawings+Using+Digitized+Graphics+Tablet
        - http://archive.ics.uci.edu/ml/datasets/Motion+Capture+Hand+Postures
        - http://archive.ics.uci.edu/ml/datasets/News+Aggregator
    - Compare 2 machine learning techniques for the modelling and prediction of data in the chosen dataset.
    - You are expected to implement these two ML techniques from first principles in the Python language. This means that you are only expected to make use of implemented third party libraries for comparative reasons.
    - In your experiments, you must make use of the following techniques
        - re-scaling and normalization
        - cross-validation
        - Principal Component Analysis
        - feature selection
        - evaluation method
    - Write and hand in a report where you explain all the steps of the implemented methodology and experiments carried out.
        - *Skeleton of the report*
            - Section 1 - Introduction
                - Explain the properties of the chosen data set and what you will be doing with it
                - Mention the two (or more) machine learning techniques that you will be using
            - Section 2 - Background
                - Describe the mechanics of the selected machine learning techniques
                - Describe what rescaling and normalisation are and why they are important
                - Describe what cross validation is
                - Describe what dimensionality reduction and feature selection methods are
                - Explain the quantitative measurements that you will be using to quantify the results; e.g. accuracy rate
            - Section 3 - Experiments
                - Describe the steps that you used to process the data set
                - Describe the experiments that you carried out
                - Describe the implementation of the 2 ML techniques chosen
                - Compare your implementation of these techniques using the dataset against a third party implementation of the same techniques.
            - Section 4 - Conclusions

            - Draw conclusions from your experiments. Example feature selection with entropy works better than PCA, ...
        - *Citations*
            - When you use citations, please use the IEEE standard referencing style.
        - *Formatting of Report*
            - It should not be longer than 20 A4 pages including everything
            - Single line spacing
            - Font type Arial, font size 10
4. *Dates:*
    - The documentation must be handed in to Ms. Francelle Scicluna (Level 1, Block A, Room 4) not later than the date stipulated above. This is a hard deadline. There will be 10% deducted with every one hour of delay.
    - Presentations as below.

10% Presentation (10% of unit marks)

- You can work in groups of 3, same team as above.
- Every group is expected to present the work done in the 80% component above.
- The presentation will take 10 minutes + 5 minutes discussion
- The date of the presentations is yet to be announced. An alphabetical order will be considered.
- The 10 points are divided as follows:
    - 50% - average of the grades anonymously collected from the audience (i.e. your fellow students)
    - 40% - quality of your presentation assessed by the lecturers
    - 10% - active participation during the discussions of other students' presentations

For each presentation, every student in the audience will fill out a Google Form. The given scores and remarks will only be visible to the lecturers and will remain anonymous to the presenters.