University of Lincoln Assessment Framework
Assessment 2 Briefing 2020-2021

Module Code & Title: CMP3751M CMP9772M-Machine Learning

Contribution to Final Module Mark: 50%

Description of Assessment Task and Purpose:

Task Overview: Cancer Classification

The objective of this assessment is to analyse a dataset comprising a number of cancer screening tests. Cancer is a condition where cells in a specific part of the body grow and reproduce uncontrollably. The cancerous cells can invade and destroy surrounding healthy tissue, including organs. 1 in 2 people will develop some form of cancer during their lifetime. In the UK, there are around 367,000 new cancer cases every year, that's around 1,000 every day (2015-2017). Cancer screening tests aim to find cancer early, before it causes symptoms and when it may be easier to treat successfully.

The dataset has two classes (healthy/cancerous condition) and 9 clinical features, including age, body mass index (BMI), glucose, insulin, leptin, resistin, MCP.1, HOMA, and adiponectin. The class membership of each row is stored in the field “Status”. Status refers to the health condition of patients, or in other words, we consider this to be our label/annotation for the sake of all implementations. Unit of measurement or range of values of each feature are not relevant. However, features can be at different scales and/or measured in different units. Our task is to develop a set of classification models for automatically classifying patients as healthy or cancerous, based on the clinical features from cancer screening. No prior knowledge of the domain problem is needed or assumed to fulfil the requirements of this assessment whatsoever.

You need to write a report that discusses how you completed the tasks and go into enough depth to demonstrate knowledge and critical understanding of the relevant processes involved. 100% of available marks are through the completion of the written report, with clear and separate marking criteria for each required report section.

Report Guidance

Your report must conform to the below structure and include the required content as outlined in each section. Information on specific marking criteria for each section is available in the accompanying CRG document. You must supply a written report containing four distinct sections that provide a full and reflective account of the processes undertaken.

Section 1: Data import, summary, pre-processing and visualisation (20%)

As a first step, you need to load the data set from the .xlsx file into a Python IDE. You should then provide a summary of the dataset (e.g. mean values, standard deviations, min/max values, etc. for each feature) and proceed to data pre-processing. For example, what is the size of the data? How many features are there? Are there any missing values? Are there any categorical variables? Shall we normalise the data before starting training/testing any model? (10%)

To visualise the data, you need to generate two plots. The first one shall be a box plot, which will include the two classes (“Status”), i.e. healthy/cancerous, in the x-axis and the “Age” in the y-axis. The second one shall be a density plot for the feature “BMI”, with the graphs of both classes appearing in the same plot. What information can I obtain from each of these two plots? Can one use any of these two plots to identify outliers? If yes, please elaborate. (10%)

Please include your explanation of implementation alongside the plots.

Hint: You can use available libraries in Python, e.g. pandas.

Section 2: Discussion on selecting an algorithm (30%)

A cancer research group is planning to use the dataset that has been provided to you to train a classifier to aid doctors in their clinical practice to detect whether a patient is developing a cancer or not. This would be very useful as clinical features/parameters are constantly monitored and stored, and therefore can be used to classify patients' health conditions. A Machine Learning intern is asked to select the best performing model among many trained models (e.g. many types of classifiers have been trained e.g. KNN, SVM, Neural Networks, etc.). The intern used 70% of the data as training set, another 20% as validation set, and finally a 10% as test set. The intern trained 10 different models (either by selecting a subset of the available features or by using a different type of classifier) and recorded the accuracy on the test set. Intern's best performing model achieves 90% of accuracy. Intern concludes that this model is the best one to use. Would you agree? Why? Please explain and elaborate.

Section 3: Designing algorithms (30%)

You will now design an artificial neural network (ANN) classifier for classifying patients as healthy or cancerous, based on their clinical features. You will use the provided data set to train the model. To design an ANN, use a fully connected neural network architecture with two hidden layers; use the sigmoid function as the non-linear activation function for the hidden layers and logistic function for the output layer; set the number of neurons in each hidden layer to 500. Now randomly choose 90% of the data as training set, and the rest 10% as test set. Train the ANN using the training data, and calculate the accuracy, i.e. the fraction of properly classified cases, using the test set. Please report how you split the data into training and test sets. In addition, please report the steps undertaken to train the ANN in detail and present the accuracy results. (12%)

You will try different numbers of epochs to monitor how accuracy changes as the algorithm keeps learning, which you can plot using the number of epochs in the 'x' axis and the accuracy in 'y' axis. (3%)

Now use the same training and test data to train a random forests classifier. Set a) number of trees = 1000 and b) minimum number of samples required to be at a leaf node = {5 and 50}. Please report the steps for training random forests and show the test set accuracy results. (12%)

You can play with tweaking the number of trees and monitor how the performance changes as more trees are added, e.g. 10, 50, 100, 1000, 5000 trees and so on. (3%)

Section 4: Model selection (20%)

You have designed ANN and random forests classifiers with almost fixed model parameters. The performance of the model could vary when those model parameters are changed. You would like to understand, which set of parameters are preferable, and also to select the best set of parameters given a range of options. To do so, one method is to employ a cross-validation (CV) process. In this task, you are asked to use a 10-fold CV. As a first step, randomly split the data into 10 folds of nearly equal size, and report the steps undertaken to do this. (8%)

For ANN, set each hidden layer (remember there are two hidden layers) with 50, 500, and 1000 neurons. For random forests, set number of trees to 20, 500, and 10000, with the "minimum number of samples required to be at a leaf node" chosen by you (or otherwise default value, if you opted to use the pre-built library and not develop it yourself; either way report what this value is). (2%)

Please do the following tasks for both methods:
a) Use the 10-fold CV method to choose the best number of neurons or number of trees for ANN and random forests respectively. b) Report the processes involved when applying CV to each combination/model. c) Report the mean accuracy results for each set of parameters, i.e. for different number of neurons and different number of trees accordingly. Which parameters should we use for each of the two methods, i.e. specifically for ANN and random forests? (5%)

b) Until now, you have selected the best parameters for each method, but we have not decided yet, which the best model is. With the results you have had so far, which one is the best model across all combinations of ANNs and random forests for this data set? Please discuss and justify your choice, reflecting upon your knowledge thus far. (5%)

Learning Outcomes Assessed:

On successful completion of this component a student will have demonstrated competence in the following areas:

- [LO1] Critique and appraise the scope and limits of machine learning methods by identifying their strengths and weaknesses
- [LO2] Using a non-trivial dataset, plan, execute and evaluate significant experimental investigations using multiple machine learning strategies

Knowledge & Skills Assessed:

Subject Specific Knowledge, Skills and Understanding: You are expected to demonstrate a good knowledge of machine learning algorithms, by exploring the machine learning literature. You will show your ability to use these algorithms to complete tasks set in this assignment. You will also demonstrate your academic writing skills.

Personal Skills: You will work independently to complete tasks. The assessment will also assess your responsibility to manage your coding and data. It will explore your creativity and problem-solving skills.

Assessment Submission Instructions:

The report should be a **maximum of 12 pages (including everything!).** Keep in mind that:

- The report must contain your name, student number, module name
- The report must be a single PDF file
- The report must be formatted in **single line spacing** and use an **11pt font**
- The report does not include this briefing document, no cover page and table of content
- You describe and justify each step that is needed to reproduce your results by using code-snippets, screenshots and plots

The deadline for submission of this work is included in the School Submission dates on Blackboard. You must make an electronic submission of your work to **Blackboard** the **Turnitin upload area** for assessment item 2.

This assessment is an individually assessed component. Your work must be presented according to the School of Computer Science guidelines for the presentation of assessed written work. Please make sure you have a clear understanding of the grading principles for this component as detailed in the accompanying Criterion Reference Grid. If you are unsure about any aspect of this assessment item, please seek the advice of the delivery team.

Date for Return of Feedback:

You can find the date for the return of feedback in the School Hand-in Schedule which is available on Blackboard under the Assessment Page of this module

Format for Assessment:

This is an individual assignment. Your work must be presented according to the Lincoln School of Computer Science formatting guidelines for the presentation of assessed written work. The final submission must have a report in **PDF** format. The source codes created for this assignment can be submitted in a **zipped** file if you wish to do so. The submission should be through the Blackboard upload area for this assessment item.

Feedback Format:

Feedback will be given to you through Blackboard. Additionally, face-to-face feedback can be obtained on request with the module staff.

Additional Information for Completion of Assessment:

The delivery team strongly recommends for you to strictly follow the following points:

- Please add sensible labels and captions to plots you will provide.
- Please describe and justify each step that is needed to reproduce your results by using code-snippets, screenshots and plots. When using screenshots or plots generated from Python, please make sure they are clearly readable with sensible axis labels
- Please use references effectively to support your arguments in the report.
- Please interpret the results of your data analysis and model developments
- Explain trends, characteristics or even outliers when you summarise and describe data
- Always refer to accuracy as performance metric when reporting the "performance" of the algorithm.
- Should you decide to use prebuilt python libraries, such as scikit-learn, rather than implementing them yourself, you will need to provide extra justification for the steps undertaken to arrive to the conclusions, and also demonstrate adequate understanding of the algorithms. Analytical approach is required to arrive to credible and justifiable solutions.

Assessment Support Information:

There will be an assignment discussion during the lectures/workshops which will help you to better understand the numerous requirements of this assignment. If you are unsure about any aspect of this assessment document, please seek advice from a member of the delivery team.

Important Information on Dishonesty & Plagiarism:

University of Lincoln Regulations define plagiarism as 'the passing off of another person's thoughts, ideas, writings or images as one's own...Examples of plagiarism include the unacknowledged use of another person's material whether in original or summary form. Plagiarism also includes the copying of another student's work'.

Collusion is defined as when a student submits work for assessment done in collaboration with another person as entirely their own work or collaborates with another student to complete work which is submitted as that other student’s work. Collusion does not apply in the case of the submission of group projects, or assessments that are intended to be produced collaboratively.

Plagiarism and collusion is a serious offence and is treated by the University as a form of academic dishonesty. Students are directed to the University Regulations for details of the procedures and penalties involved.

For further information, see www.plagiarism.org