School of Computer Science

COMP47460: Introduction to Machine Learning
Semester I

Assignment 1
On: 20/10/2021
Due: 31/10/2021 18:00
Total Points: 100

- *Submit your assignment as one PDF file (not a DOC/DOCX/ODT/ZIP file) via the module Brightspace page.*
- *Include your full name and student ID number on the PDF.*
- *This assignment should be completed individually. Any evidence of plagiarism will be reported to the CS plagiarism committee and it can result in a Fail grade.*
- *Download the file* `heartdisease_<student_number>.arff` *from Brightspace. So, if your student number is* `12345678`*, then download* `heartdisease_12345678.arff`*. Please ensure that you are using your personalised dataset.*
- *Your objective is use the methods you have learned in the module to make predictions about heart failure.*
- *You should use Weka for this assignment. It's ok to include screenshots of the Weka output where appropriate. You can also use python (if you do, include some information on the classes/methods you used to arrive at the results).*
- *This is an open-ended assignment -- You do not have to restrict yourself to what is asked. You can take the exploration and the discussion deeper than what is asked.*
- *Total suggested page length is around 5 pages.*
- *Concise is better than effusive.*

Q1: (**100points**)

You will analyse a dataset relating to heart failure prediction.

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, estimated to be responsible for 17.9 million lives each year. This accounts for 31% of all deaths worldwide. Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

The features included in the dataset are:

1. **Age**: age of the patient [years]
2. **Sex**: sex of the patient [M: Male, F: Female]
3. **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. **RestingBP**: resting blood pressure [mm Hg]
5. **Cholesterol**: serum cholesterol [mm/dl]
6. **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. **RestingECG**: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
9. **ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]
10. **Oldpeak**: oldpeak = ST [Numeric value measured in depression]
11. **ST_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. **HeartDisease**: output class [1: heart disease, 0: Normal]

1.1. Examine the dataset carefully in the Weka Explorer. You should normalise and/or clean the dataset, as appropriate. Describe the data preparation/cleaning steps you took (e.g., Your description may look like "I did min-max normalisation of feature X using the minimum on the feature values in training examples and maximum of feature values over all labelled examples. I manually removed feature Y because ..."). [5]

1.2. Use 3 classifiers from the module and compare the accuracy of the **HeartDisease** predictions. You should choose the accuracy measures, explain your choices, and discuss some reasons for the different accuracy values. [10]

1.3. This dataset has many features. Carefully identify the most discriminating features to predict **HeartDisease** using the filter and wrapper feature selection techniques. [25]

i. Report the feature subsets that these techniques select. In the case of a filter, you should propose a way to choose a subset of the ranked features, rather than using the entire original set of features, and justify your choice. In the case of wrapper techniques, carefully select features for at least one Decision Tree, one Naïve Bayes and one k-NN classifier.

ii. Report and discuss the differences between the feature subsets produced by the filter and wrapper techniques.

iii. Carefully consider the evaluation measure(s) that you use for this exercise and justify why you selected the particular evaluation measure(s).

1.4. Evaluate the performance of various classifiers (including at least one Decision Tree, one Naïve Bayes and one k-NN classifier) on your dataset using the feature subset(s) identified in 1.3.i. and evaluation measures identified in 1.3.iii. Explore the effect of different parameter settings on these classifiers. Describe the evaluation procedure that you used in detail. [20]

1.5. To what extent are these results in line with or different from what you learnt about these classifiers in your lectures? For example, is your accuracy higher or lower on the dataset with reduced number of features as compared to the original dataset? Is the relative performance of different classifiers and configuration settings in line with your expectation? [20]

1.6. [20]

i. Plot the ROC curves for 3 different classification models (Weka does this).

ii. What do you learn from these ROC curve? Include the AUC in your discussion.

iii. Which classifier/configuration is best suited for this task? Are you satisfied with the performance?