School of Computer Science

COMP47460: Introduction to Machine Learning
Semester I

Assignment 2
On: 17/11/2021
Due: 28/11/2021 23:59
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 `tesco_<student_number>.arff` from Brightspace. So, if your student number is `12345678`, then download `tesco_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 collected from shoppers at Tesco stores in the UK. The dataset has a number of features describing aggregated purchasing behaviour in different geographic locations. These features relate to the quantity of items purchased and their nutritional information, in addition to other features.

Researchers have used this dataset to try to predict the prevalence of diabetes, and you might find the paper to be an interesting read. Some information on the features in the dataset can also be found on Figshare.

The last column in the dataset is a categorical feature describing the diabetes prevalence {low, mid, high}.

The objective of this question is to use the ensemble learning functionality to identify the extent to which classification performance can be improved through the combination of multiple models.

1.1. Evaluate the performance of three basic classifiers on your dataset: ☒Decision Tree, Neural Network and 1-NN. Carefully consider the evaluation measure(s) that you use for this exercise and justify why you selected the particular evaluation measure(s). [10]

1.2. Use the Weka Vote ensemble method (meta -> Vote) to combine the Decision Tree, Neural Network and 1-NN classifiers. Evaluate the performance of the Vote ensemble method with 3 different combination rules (there are 6 possibilities: Average of probabilities, Product of probabilities, Majority voting, Minimum probability, Maximum probability, Median). Provide a justification for the difference in accuracy when using different combination rules. [15]

1.3. Some of the features may be correlated with others or have dependencies on other features. Build a linear regression model to predict the carb feature from this set of features: f_beer, f_dairy, f_eggs, f_fats_oils, f_fish, f_fruit_veg, f_grains, f_meat_red, f_poultry, f_readymade, f_sauces, f_soft_drinks, f_spirits, f_sweets, f_tea_coffee, f_water, f_wine. [15]

Weka chooses the feature which minimises the residual error- which feature does it find? Show the regression model and comment on the quality of the model.

1.4. Return to the full data set and apply ensembles with bagging using the three classifiers from Task (a). Investigate the performance of these classifiers as the ensemble size increases (e.g., in steps of 2 from 2 to 20 members). Using the best performing ensemble size, investigate how changing the number of instances in the bootstrap samples affects classification performance (i.e. the "bag size"). [25]

1.5. Apply ensembles with random subspacing using the three classifiers from Task (a). Investigate the performance of these classifiers as the ensemble size increases (e.g., in steps of 2 from 2 to 20 members). Using the best performing ensemble size, investigate how changing the number of features used when applying random subspacing affects classification performance (i.e. the "subspace size"). [25]

1.6. Based on the lectures, which set of classifiers is expected to benefit from bagging techniques more and which set of classifiers is expected to benefit from random subspacing techniques more? For your dataset, determine the best ensemble strategy for each of these classifiers. Discuss if this is in line with what you expected. [10]