COMP9417 20T1 Homework 2

Applying and Implementing Machine Learning

Introduction

The aim of this homework is to enable you to:

1. **apply** parameter search for machine learning algorithms implemented in the Python *sklearn* machine learning library
2. **answer questions** based on your **analysis** and **interpretation** of the empirical results of such applications, using your knowledge of machine learning
3. **complete an implementation** of a different version of a learning algorithm you have previously seen

After completing this homework, you will be able to:

- set up a simple grid search over different hyper-parameter settings based on k-fold cross-validation to obtain performance measures on different datasets
- compare the performance measures of different algorithm settings
- propose properties of algorithms and their hyper-parameters, or datasets, which may lead to performance differences being observed
- suggest reasons for actual observed performance differences in terms of properties of algorithms, parameter settings or datasets.
- read and understand incomplete code for a learning algorithm to the point of being able to complete the implementation and run it successfully on a dataset.

There is a total of *5 marks* available.

Deadline: **Monday 30 March 2020, 5:00 PM**

Submission will be via the Moodle page.

Late penalties: one mark will be deducted from the total for each day late, up to a total of five days. If six or more days late, no marks will be given.

Recall the guidance regarding plagiarism in the course introduction: this applies to this homework and if evidence of plagiarism is detected it may result in penalties ranging from loss of marks to suspension.

There are two parts of questions in this homework, for the first part you should run a python code using provided dataset and answer some questions related to that. For second question, you should build Decision Tree model using titanic dataset.

Question 1

In this question, you should run the 'comp9417_hw2.ipynb' and the provided results will help you to find the answer of the following questions.

Dataset

You can download the datasets required for the homework that are included in the homework zip file on Moodle. **Note:** you will need to ensure the dataset files are in the same directory from which you are running this notebook.

Package installation

Please Note: this homework uses some datasets in the Attribute-Relation File Format (.arff). To load datasets from '.arff' formatted files, you will need to have installed the "scipy.io" package. You can do this using pip at the command-line with the following code, if this package is not installed yet:

```
pip install scipy
```

Part A - [0.5 marks]

The Python code in the notebook 'comp9417_hw2.ipynb', tests Decision Tree and Naïve Bayes methods by training them on variable training data size, from 5% to 50%. A **learning curve** is observed if the accuracy has a general tendency to increase with a larger training size.

To answer this question, you should run the python code in the notebook 'comp9417_hw2.ipynb'. You will have two tables which you should copy and paste into your report as your answer for "Question Part A". These tables are similar to the below:

Copy/paste the output from running the code in the notebook. DO NOT EDIT the table!!!

DecisionTreeClassifier

Dataset	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
australian	79.86%	81.29%	82.91%	82.02%	82.17%	81.45%	82.03%	83.34%	83.33%	82.33%
balance-scale	75.67%	75.99%	76.98%	77.92%	77.30%	78.10%	77.95%	77.62%	78.09%	77.45%
hypothyroid	99.42%	99.52%	99.20%	99.28%	99.23%	99.31%	99.34%	99.52%	99.52%	99.47%

BernoulliNB with priors

Dataset	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
australian	79.84%	81.14%	81.28%	81.29%	82.74%	82.16%	82.16%	82.74%	82.74%	82.74%
balance-scale	46.08%	46.08%	46.08%	46.08%	46.08%	46.08%	46.08%	46.08%	46.08%	46.08%
hypothyroid	92.26%	92.23%	92.23%	92.18%	92.21%	92.23%	92.26%	92.26%	92.26%	92.23%

Part B - [0.5 marks]

Looking at the results for BOTH the Bernoulli Naive Bayes and Decision Tree learning models over all the datasets, the following is true regarding a possible "learning curve" effect due to increasing the size of the training set (choose ALL true statements):
--- (1) none of the 6 models show a learning curve
--- (2) all of the 6 models show a learning curve
--- (3) most of the 6 models show a learning curve
--- (4) All 3 Decision Tree models are generally better than Bernoulli Naive Bayes models
--- (5) Some Bernoulli Naive Bayes models are better than Decision Tree models

Part C - [0.5 mark]

Bernoulli Naive Bayes (BNB) in the code used in the previous question use class priors as a default. In some cases better results can be obtained without using priors.
In this question you will find how the results are affected by removing class priors (in this case replacing them with uniform probabilities). After the last line, add one more line of code for BNB but with uniform priors and re-run the code. Consult sklearn documentation which option to use.
After adding the new line, compare BNB model results with and without priors and choose ONE true statement from the below:
--- (1) BNB preforms better with priors
--- (2) BNB preforms better without priors
--- (3) there is no difference in performance when using BNB with or without priors

Question 2

In this practical question, you will develop a machine learning model using Decision Tree to predict which passenger will survive the Titanic sinking. You will also calculate some probabilities based on your knowledge from lectures and tutorials. In order to do this, you will use “titanic.csv” file with 887 instances. For easy pre-processing, the column values are coded as integers with the following meaning:
'Pclass': passenger class: 1, 2 or 3
'Sex': 0 male, 1 female
'Age': 0: 0-9 years old, 10: 10-19 years old, and so on
'Siblings_Spouses_Aboard': number of siblings and/or spouses traveling with a passenger
'Parents_Children_Aboard': number of parents (if the passenger is a child) or number children (if the passenger is a parent)
'Survived': 0 did not survived, 1 survived
Then you should be able to find the answer for the questions.

Pre-processing

In the first step of this homework, you can start by applying min-max normalisation to all features (). You can test whether you did the normalisation correctly or not by checking the minimum and maximum value for each of the features of the whole dataset.

Min-Max normalization

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

After applying this normalisation, the minimum value of your feature will be 0 and the maximum value will be 1. You are also allowed to use built in functions to do the normalisation.

Creating test and training sets

In this step, you have to create training and test sets. Please use the first 620 rows of the data (70%) as training set and keep the 267 remaining rows (from 620 to 887) as test set which you will use later to evaluate the model.

Part A - [1 mark]

Implement a Decision Tree Classifier for survival prediction using sklearn library with all other parameters left default. What is your accuracy score for training and test dataset?

Part B - [1 mark]

Find an optimal number of min_samples_leaf by developing a search algorithm to search min_samples_leaf between 2 and 20. Please use AUC score in the search.

Part C - [0.5 mark]

Plot the AUC score for all iterations (k: 2,…,20) in training and test sets. (one plot for training, and one for test set).

Part D - [1 mark]

Using the same titanic dataset, calculate the posterior probability that a female passenger, who travels in the first class, survives. Use full titanic dataset to calculate this probability.

P(S= true | G=female,C=1)

!!!!You are required to copy and paste all your code at the end of your report, AND to submit the Python file(s) with all your code as a separate file as well!!!!