COMP9417 19T3 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: **Friday 1 November 2019, 5 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 k-nearest neighbour model using Australian credit card 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 in the datasets folder. **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]

To answer this question, you should run the python code in the notebook 'comp9417_hw2.ipynb'. you will have a table similar to the bellow table which you should copy and paste in your report as your answer for "Question Part A".

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

Decision Tree Results

Dataset	Default	0%	25%	50%	75%
australian	35.11% (2)	77.66% (17)	72.34% (22)	55.85% (27)	20.21% (7)
labor	25.49% (2)	47.06% (12)	42.16% (7)	32.35% (12)	18.63% (2)
diabetes	21.54% (2)	47.69% (2)	43.08% (7)	49.23% (12)	21.54% (2)
ionosphere	47.19% (2)	76.40% (22)	76.40% (22)	47.19% (2)	22.47% (7)

Part B - [0.5 marks]

In sklearn's decision tree implementation, the parameter "max_depth" represents the maximum depth of decision tree during the training procedure. The deeper the tree, the more splits it has, and it captures more information about the data.

By increasing the value of the value of "max_depth" parameter we can expect this to:
--- (1) overfitting not changed by decreasing max_depth of the decision tree
--- (2) decrease overfitting by increasing max_depth of the decision tree
--- (3) increase overfitting by decreasing max_depth of the decision tree
--- (4) increase overfitting by increasing max_depth of the decision tree

Part C - [0.5 mark]

Looking at your table, the performance result for datasets with 50% noise and "max_depth" parameter. Does finding the best parameter in grid search helps the decision tree model to improve the test set accuracy compared to the default parameter settings? What is your answer?

--- (1) no
--- (2) yes, for 1/4 of the datasets
--- (3) yes, for 2/4 of the datasets
--- (4) yes, for 3/4 of the datasets
--- (5) yes, for 4/4 of the datasets

Question 2

In this practical question, you should develop a machine learning model using K-nearest neighbours to predict which customer will be able to return the loan without any delinquency. In order to build your model, you should use CreditCard.csv file in Datasets folder. Then you should find the answer of the following questions:

Pre-processing

In the first step of this homework, you can start by applying min-max normalisation to all features (X1 – X14). You can test whether you did the normalisation correctly or not by checking the minimum and maximum value for each of your features.

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.

Creating test and training sets

In this step, you have to create training and test sets. Please use the first 621 rows of the data as training set and keep the 69 remaining one (from 622 to 690) as test set which we will use later to evaluate the kNN model.

!!!!Please copy and paste all your codes in your submission file!!!!

Part A - [1 mark]

Implement a kNN classifier for Australian credit risk prediction using sklearn library. You should set the n_neighbors =2 for training the model. What is your accuracy score for training and test dataset?

Part B - [1 mark]

Find optimal number of neighbours by developing a search algorithm to find the optimal value of k. You should find the optimal number of k in a range between 1 to 30 and finding optimal value for number of k. please use AUC score to find the optimal number of neighbours.

Part C - [0.5 mark]

Plot the AUC score for all iterations (k: 1,…,30) in training and test sets. (one plot for training, and one for test set).

Part D - [1 mark]

Compute precision and recall evaluation metrics for your kNN model with optimal number of neighbours and another model that you have built in part A. Compare these metrics for these two models.