F20/21DL Data Mining and Machine Learning: Coursework Assignment 2

Handed Out: 2nd November 2020
What must be submitted: a report of maximum 4 sides of A4 (5 sides of A4 for Level 11), in PDF format, and accompanying software.
Submission deadline: 15:00pm Monday 30th November 2020 -- via Vision
Worth: 25% of the marks for the module**.**

<u>The point</u>: this coursework is designed to give you experience with, and hence improve your understanding of:

- Overfitting: finding a classifier that does very well on your training data doesn't mean it will do well on unseen (test) data.
- The relationship between overfitting and complexity of the classifier – the more degrees of freedom in your classifier, the more chances it has to overfit the training data.
- The relationship between overfitting and the size of the training set.
- Bespoke machine learning: you don't have to just use one of the standard types of classifier – the application may specifically require a certain type of classifier, and you can develop algorithms that find the best possible such classifier.

The data set:

The data set for the coursework is the same as in Coursework 1, please refer to Coursework 1 specification for general description and motivation. For this coursework, we additionally provide the testing data sets. They can be downloaded here: http://www.macs.hw.ac.uk/~ek19/data2/. The naming convention is as follows:

1. The main training data set: as in coursework 1.
2. The main test data set:

- [x_**test_gr_smpl.csv**] contains test features (i.e. the images).

3. Class labels for the main test data set:

- **[y_test_smpl.csv]** contains labels for the test file, ranging from 0 to 9.
- [y_**test_smpl_<label>.csv**] Test labels for one-vs-rest classification. Images that were originally in class <label> are marked by 0 and all other images -- by 1. For example, if <label> is 6, then all images in the train set displaying a stop sign are given the label 0, labels of all other images are set to 1.
- Arff versions of the above files are given.
- *Note 1: the ``one-vs-rest" data sets can be used as reserve data sets, for testing various hypotheses you may come up with in the coursework. Also they may give better accuracies, and thus may be handy for some experiments. Please use them to enrich your research hypotheses and experiments.*

What to do:

<u>Before you start:</u> Choose the software in which to conduct the project: we recommend Python for the part concerning Neural Networks. Create folders on your computer to store classifiers, screenshots and results of all your experiments, as explained below.

Your experiments will consist of two parts – in Part-1 you will work with Decision trees and in Part -2 – with Linear Classifiers and Neural Networks.
For each of the two parts, you will do the following:

1. Using the provided training data sets, and the 10-fold cross validation, run the classifier, and note its accuracy for varying learning parameters. Record all your findings and explain them. Make sure you understand and can explain logically the meaning of the confusion matrix, as well as the information

contained in the "Detailed Accuracy" field: TP Rate, FP Rate, Precision, Recall, F Measure, ROC Area.

2. Use Visualization tools to analyze and understand the results.
3. Repeat steps 1 and 2, this time using training and testing data sets instead of the cross validation. That is, build the classifier using the training data set, and test the classifier using the provided test data set. Note the accuracy.
4. Make new training and testing sets, by moving 4000 of the instances from the original training set into the testing set. Then, repeat step 3.
5. Make new training and testing sets again, this time removing 9000 instances from the original training set and placing them into the testing set again repeat step 3.
6. Analyze your results from the point of view of the problem of classifier over-fitting.

Level 11 only (MSc students and MEng final year students):

7. *[Research Question]* Think about your own research question and/or research problem that may be raised in relation to the given data set, and the topics of Decision Tree learning, Linear Classifiers and Neural Networks. Formulate this question/problem clearly, explain why it is of research value. The problem may be of engineering nature (e.g. how to improve automation or speed of the algorithms), or it may be of exploratory nature (e.g. something about finding interesting properties in data), -- the choice is yours.
8. *[Answer your research question]* Provide a full or preliminary/prototype solution to the problem or question that you have posed. Give logical and technical explanation why your solution is valid and useful.

Detailed technical instructions:

Part 1. Decision tree learning.

In this part, you are asked to explore decision tree algorithms:

1. J48 Algorithm
2. One other Decision tree algorithm of your choice (e.g. random forest).

You should compare their relative performance on the given data set. For this:

- Experiment with various decision tree parameters: binary splits or multiple branching, pruning, confidence threshold for pruning, and the minimal number of instances permissible per leaf.
- Experiment with their relative performance based on the output of confusion matrices as well as other metrics (TP Rate, FP Rate, Precision, Recall, F Measure, ROC Area). Note that different algorithms can perform differently on various metrics. Does it happen in your experiments? – Discuss.
- Record all the above results by going through the steps 1-6.

Part 2. Neural Networks.

In this part, you will work with Neural Networks:

- Run a Linear classifier on the data set. This will be your base for comparison.
- Run a *Multilayer Perceptron*, experiment with various Neural Network parameters: add or remove layers, change their sizes, vary the learning rate, epochs and momentum, and validation threshold.
- Experiment with relative performance of Neural Networks and changing parameters. Base your comparative study on the output of confusion matrices as well as other metrics (TP Rate, FP Rate, Precision, Recall, F Measure, ROC Area).
- Record all the above results by going through the steps 1-6.
- For higher marks, try running *Convolutional Neural Networks*, and repeat all of the above steps for them.

What to Submit:

You will submit:

a) All sources with the evidence of conducted experiments: data sets, scripts, tables comparing the accuracy, screenshots, etc. Give a web link to them (Github, Bitbucket, Dropbox, own webpage…).

b) A report of maximum FIVE sides of A4 (11 pt font, margins 2cm on all sides) for Honours BSc students and SIX sides of A4 (11 pt font, margins 2cm on all sides) for MSc students. Figures and tables do not count towards the page limit. Figures and illustrations do not count as part of the page limit. Only one report per group should be submitted, as multiple submissions will trigger the plagiarism checker.

c) Declaration of what parts of the coursework each group member contributed. Data preparation, programming, analysis, report writing (and generation of figures and illustrations for the report) all count as contribution. You are encouraged to solve more complex research tasks collaboratively: you will learn more and do a better job if you discuss your progress and your actions regularly.

Using the results and screenshots you recorded when completing the steps 3-8, write five sections, respectively, entitled:

1. **"Variation in performance with size of the training and testing sets"**
2. **"Variation in performance with change in the learning paradigm (Decision Trees versus Neural Nets)"**
3. **"Variation in performance with varying learning parameters in Decision Trees"**
4. **"Variation in performance with varying learning parameters in Neural Networks"**
5. **"Variation in performance according to different metrics (TP Rate, FP Rate, Precision, Recall, F Measure, ROC Area)"**
6. **(Level 11 students) Own research topic.**

In each of these sections you will speculate on the reasons that might underpin the performance and the variations that you see, considering general issues and also issues pertaining to this specific task. You are recommended to represent all your results in figures or tables – to which you will refer from these five specific sections.

Marking:

Points possible: 100.

Level 10: Each Section is worth 20 points of the total 100 points.
Level 11: **Sections 1-5 are worth 17 points each, section 6 is worth 15 points.**

The exact marking rubrics are available on Vision.

Plagiarism

This project is assessed as **group work**. You must work within your group and not share work with other groups. Readings, web sources and any other material that you use from sources other than lecture material must be appropriately acknowledged and referenced. Plagiarism in any part of your report will result in referral to the disciplinary committee, which may lead to you losing all marks for this coursework and may have further implications on your degree. https://www.hw.ac.uk/students/studies/examinations/plagiarism.htm

Lateness penalties

Standard university rules and penalties for late coursework submission will apply to all coursework submissions. See the student handbook.