F21DL Data Mining and Machine Learning: Coursework Assignment 3
Handed Out: November 7th 2016
What must be submitted: A report of maximum THREE sides of A4, in PDF format
To be 'Handed in': 23:59pm Sunday December 4th 2016 - via Vision
Worth: 30% of the marks for the module**.**

The point: this coursework is designed to give you experience with, and hence more 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'client' may specifically want a certain type of classifier (here, a ruleset that works in a certain way), and you can develop algorithms that try to find the best possible such classifier.

In this coursework you will work with only the "Optical recognition of handwritten digits" dataset from the UCI repository. You will use the following specific training and testing sets, available on Vision:
optraining.txt,
optesting.txt.

The raw data of the 'optical recognition' dataset came from 32x32 pixellated images of handwritten digits, (e.g. either '0', '1', '2', etc…, '9') . where each pixel was either black or white. These were preprocessed to become 8x8 arrays, where each pixel represented a 4x4 square from the raw data, and hence is a number between 0 and 16 inclusive. This is the data you see in the above files. Each instance gives the 64 pixels in row by row order, with the last field (the 65th field) indicating the digit that is represented in this image.

What to do (an overview):

Your coursework will consist of two parts – in Part-1 you will work with Decision trees and in Part -2 – with neural networks.

Before you start:
1. Convert the above files into arff format and load them to Weka.
2. Create folders on your computer to store classifiers, screenshots and results of all your experiments, as explained below.
As part of your coursework marking, you may be asked to re-run all your experiments in the lab. So please store all of this data safely in a way that will allow you to re-produce your results on request.

For each of the two parts (Decision Trees and Neural Networks), you will do the following:
3. Using the provided data set, and Weka's facility for 10-fold cross validation, run the classifier, and note its accuracy for varying learning parameters provided by Weka. (Below you will find more instructions on those.) 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.
4. Use Visualization tools to analyze and understand the results: Weka has comprehensive tools for visualization of, and manipulation with, Decision trees and Neural Networks.
5. Repeat steps 3 and 4, this time using testing data set instead of Weka's cross validation.
6. Make new training and testing sets, by moving 900 of the instances in 'optesting.txt' into the training set. Then, repeat steps 3 and 4.
7. Make new training and testing sets again, this time with 2000 instances in the training set, and the remainder in the test set, and again repeat steps 3 and 4.
8. Analyse your results from the point of view of the problem of classifier over-fitting.

Detailed technical instructions:

Part 1. Decision tree learning.

In this part, you will be asked to explore the following three decision tree algorithms implemented in Weka

1. J48 Algorithm
2. User Classifier (This option allows you to construct decision trees semi-manually)
3. One other Decision tree algorithm.

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

- Experiment with various decision tree parameters: binary splits or multiple branching, prunning, 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.
- When working with User Classifier, you will need to learn to work with both Data and Tree Visualizers in Weka.
- Record all the above results by going through the steps 3-8.

Part 2. Neural Networks.

In this part, you will work only with the *MultilayerPerceptron* algorithm in Weka.

- Experiment with various Neural Network parameters: add or remove nodes, layers and connections, vary the learning rate, epochs and momentum, and validation threshold.
- You will need to work with Weka's Neural Network GUI in order to perform some of the above tasks efficiently.
- Experiment with the 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 3-8.

What to Submit

You will submit a report of maximum THREE sides of A4 (11 pt font, margins 2cm on all sides), containing the following.

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

- "Variation in performance with size of the training set"
- "Variation in performance with change in the learning paradigm (Decision trees versus Neural Nets)"
- "Variation in performance with varying learning parameters in Decision Trees"
- "Variation in performance with varying learning parameters in Neural Networks"
- "Variation in performance according to different metrics (TP Rate, FP Rate, Precision, Recall, F Measure, ROC Area)"

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

The material must be all contained within 4 sides of A4. Each section should take up to half of a page, the table and screenshots should take up to 1.5 page.
20 marks are lost for every extra page, even if there is just one word on the page.

Marking: CW3 is worth 30% of the module; of that 30%. The above parts break down as follows:

Each Section is worth 20% of the total 100% of CW3 mark.
For top marks, you are expected to show deep understanding of all of the involved machine learning techniques and methods, mastery of the Weka toolbox, as well as ability to do data mining and machine learning research, i.e. to analyse research outputs clearly, logically, soundly and convincingly.

The minimum pass mark will be given to submissions in which steps 3-7 are conducted for Decision trees (at least one algorithms) and Neural Networks, all of the results are clearly recorded, but the author fails to gather, analyze and critique the results or to show deep understanding of the methods and tools involved in the task.