CS36110: Machine Learning – Assignment 2

THE MNIST DATABASE of handwritten digits

Release date: Thursday 10th November 2016
Hand in date: Wed, 14th December 2016 (18:00 via Blackboard/Turnitin)
Feedback date: Friday, 13th January 2017

This is the second of two assignments for CS36110, and comprises 60% of the total mark for the module. It will be marked out of 100% and be assessed according to the Department's assessment criteria for essays. In particular, marks will take into account the understanding of the problem, completion of the task and quality of the presentation. Other marks will cover knowledge of the lecture material, justification for the choices made, and quality of analysis.

Introduction

In this assignment, we will look at the MNIST database of handwritten digits (see Figure 1), a widely used real-world data benchmark in machine learning which is available at

`http://yann.lecun.com/exdb/mnist/`.

Figure 1: Example

You should use the .arff files provided on Blackboard for your experimentation.

The dataset contains 60 000 training samples and 10 000 test examples of handwritten digits, with roughly the same number of samples for each of the 10 different digit classes. It is a subset of a larger set available from NIST[1]. For the purpose of this assignment we have created a number of different .arff files that can be used directly in WEKA[2].

The goal is to correctly classify the digits in a given image. For this task, the digits have been size-normalised and centred in a fixed-size image of 28×28 pixels (by using the centre of mass of the pixels). Each pixel corresponds to a feature and takes a grayscale value in $\{0, 1, \ldots, 255\}$ (which are sometimes mapped onto the interval $[0, 1]$). A value of 0 corresponds to a white pixel; a value of 255 to a black pixel. There are no missing values. You can find more information about the properties of the dataset and how it was constructed on the above website.

On Blackboard, you will find two versions of the data set: MNIST_784 and MNIST_576. You will need MNIST_784 for Task 1 and 2. MNIST_576 is only used in Task 2. More details are given below. Please make sure that you use the correct files as test data and training data (these are different!).

Please note that these are huge data set and that depending on the machine and machine learning method you use, learning the model based on the full data set may take considerable time and may not be manageable[3]. We have therefore created a number of smaller datasets by selecting a random subset of the images. These sets contain 1 500 and 6 000 training and 1 000 test samples, respectively. Please let us know in case you encounter any problems with the datasets provided.

You can use these sets for your experimentation where appropriate, e. g., you could start with some preliminary experiments to investigate parameter settings on smaller sets and increase the problem size gradually as appropriate. You should always clearly say what set you use and justify why do so.

[1] The National Institute of Standards and Technology, `https://www.nist.gov`

[2] WEKA: Waikato-Environment-for-Knowledge-Analysis, `http://www.cs.waikato.ac.nz/ml/weka/`

[3] On my computer, I was able to learn a model for the full dataset with NaiveBayes in under 6 seconds and with J48 in under 4 minutes, but libSVM and MultiplePerceptron may take considerably longer depending on your choice of parameters, e. g., simply using the default parameters on the 1 500 instance took just under an hour with MultiplePerceptron.

General Guidelines

As in Assignment 1 you are asked to run experiments using WEKA (Book Version 3.6.10). If you are working on a machine outside of the departmental network (e. g. your laptop), please ensure that you download and install the same exact version. This is very important as your results need to be reproducible on a departmental machine. Moreover, we have encountered memory issues in some WEKA versions that may negatively effect your work on the assignment.

In many cases, WEKA Explorer allows you to modify the random seed that will be used. However, using the default seed is fine. If you do change the seed you need to report the seed you have chosen. As before, you will have to work out some details in applying WEKA Explorer, including definitions of certain terminologies.

Tasks

The MNIST database provides separate training and test sets that you should use to train and evaluate the classifiers. All datasets (.arff files) necessary for this assignment are available on Blackboard and can directly be used in WEKA. Start the WEKA explorer. In the "Preprocess" tab load one of the training sets provided. You can load a test set in the "Classify" tab by changing the test options to "Supplied test set" and selecting the file.

The assignment is split into two basic parts aiming at gaining some insights into the MNIST dataset and the machine learning techniques discussed in the lectures.

1. Motivated by the large number of features in the given dataset, we first look into a problem called *feature selection*, the process of selecting a subset of relevant features to be used. You are asked to perform this task by carefully considering properties of the original features and insights you gain through running a machine learning technique on the original dataset. (40%)

   You should only use the datasets labeled MNIST_748 for this part of the assignment.

   (a) In the "Preprocess" tab of the WEKA explorer you are given properties of the different features: minimum, maximum, mean, and standard deviation. The visualisation gives you additional information about the features in relation to the digit classes. Discuss how you can use this information to identify features that may be less useful for the given machine learning task than others. You should support your discussion by providing supporting data and figures. If you wish, you can also refer to properties of the dataset that you find in other resources, but you need to clearly identify your sources. (15/40)

   (b) Train and evaluate Naive Bayes (NaiveBayes) and C4.5 Decision Tree (J48) on the dataset and examine the learned classifier model (shown above the evaluation results in the output window). Discuss what you may learn about the importance of different features from these models. (15/40)

   (c) Modify the original dataset by selecting a subset of features based on your above insights. You should do this manually and not use any automated feature selection method you find under the "Select attributes" tab in WEKA. Train NaiveBayes and J48 on the new dataset and compare the results with the results for the original sets, both with respect to classification accuracy and runtime, and against a reasonable baseline. (10/40)

2. Second, you are required to train and evaluate a third classifier on the dataset. You can choose either an artificial neural network *or* a support vector machine and should briefly explain your choice in your report. You are free to perform experiments for both classifiers, but only one will be taken into account when marking your assignment. (40%)

The Neural Network implementation in WEKA is called "MultiplePerceptron" and can be found under classifiers → function. An implementation for Support Vector Machines can be found in classifiers → functions → libSVM. It should be noted that there are other SVM implementations in WEKA that provide less opportunities for customisation. If you use any of those instead, make sure that you provide sufficient information to justify your choice.

To make the results comparable we have provided a reduced version of the dataset with 576 features. You should perform your experiments on both, the original version (MNIST_748) and the reduced one (MNIST_576).

(a) Train and evaluate the chosen classifier on the two different versions of the data set and investigate different parameters and performance. Explain how you choose different parameters and the results you obtain. (20/40)

(b) Compare the performance you achieve with the performance of the NaiveBayes and J48 classifiers from the previous task. What can you say about the differences? (20/40)

Submission and Marking

You are required to submit the following elements:

1. A report in PDF format via Turnitin.
   There is no strict word limit, but 1,500 (± 10%) words (**excluding** references, footnotes, figures, tables or data in appendices) are appropriate for this assignment. Please state the word count at the beginning of your report.

2. The modified dataset created in Task 1 via Blackboard.

Your assignment will be assessed according to the department's assessment criteria for essays (see Student Handbook Appendix AC) and marked based on your report and any supporting data you submit. In general it is advisable to concentrate on appropriately sized selections or excerpts to support your discussion. It is not necessary to include very large chunks of data. However, if you feel the need to include such elements please do not include them into the main text (but in an appendix) in order not to impede reading your report. The following criteria will be used for marking your report:

- Feature Selection (40%): see description above
- Classification (40%): see description above
- Report (20%):
  - This includes formal aspects such as submission format (PDF) as well as readability, correct formatting, layout and proper referencing. All figures, tables and diagrams should include captions, and be numbered and cited where appropriate in the main text. (10/20)
  - To conclude you should discuss advantages and disadvantages of the three different classifiers. What are the best results you were able to achieve and how? Which classifier do you consider most suitable? What have you learnt about the dataset that may help you improve your results further? (10/20)

Plagiarism

One of the dangers of this assignment is the temptation to use paragraphs from web documents or papers that you have read. Please resist this temptation and do not do this. Otherwise, you will be heavily penalised. The report should be completely in your own words. If it is appropriate and absolutely necessary to include sentences and materials from elsewhere, then they should be clearly indicated as quotes, and references should be cited. Please do not show your report to other students.

Marking Grid

	70-100%	60-70%	50-60%	40-50%	0-40%
Feature Selection	Comprehensive and convincing	Very good, showing some insight.	Good, giving explanations	Basic, restatement of results, no dataset submitted	Largely missing
Classifier	Comprehensive and convincing	Very good, showing some insight.	Good, giving explanations	Basic, restatement of results	Largely missing
Report	Very well presented (e. g. using carefully selected figures and/or tables), clear and detailed description and conclusion, good use of English, good use of references.	Well presented (e. g. using some figures and/or tables), mostly clear and detailed description and conclusion, mostly good use of English, good use of references.	Presentation that may be lacking in some detail, reasonable use of English, some references.	Presentation that may have gaps or lack detail or lack appropriate support of discussion. Some poor use of English and references may be limited or missing.	Incomplete, incorrect formatting, poor use of English, no references