Machine Learning Exercises 6: Practical Assignment

Due: December 13 (has been extended by one week)

Rogier van het Schip Tim van Erven

November 30, 2007

Abstract

This practical is intended to give you hands-on experience in using a practical machine learning tool and analysing scientific results. You will first use the machine learning tool *Weka* to compare the performance of the ID3, Naive Bayes and k-Nearest Neighbour algorithms on a given dataset, and then write a report about your findings according to scientific standards.

You are allowed to work together in groups of at most *three* people and hand in a single report. There is a very strict limit on the number of pages in your report. For assistance, please contact Rogier.

1 Preliminaries

1. Form a group of at most three people.
2. Download and install Weka as explained on `http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html`.
3. You will use the data set splice.arff, which is available from the Weka website. For your convenience we have already split it into a train set, a validation set and a test set, which you should download from `http://www.cwi.nl/~erven/teaching/0708/ml/wekadata/`. (N.B. This means that contrary to what we told you in class you won't have to use cross-validation.)

2 Introduction

The data set you have downloaded contains classification data. You will have to write a report that compares the performance on this data set of three classification algorithms that have been discussed in class: ID3, Naive Bayes and k-Nearest Neighbour. The report should answer the following *main questions*:

1. Is there a difference in performance between the algorithms? Which algorithm performs best?
2. What are the reasons for the (lack of a) difference in performance?
3. Which value of k for the k-Nearest Neighbour algorithm works best (if it makes any difference)?
4. Why?

The remainder of this practical consists of instructions and smaller questions (please incorporate their answers in your report as well) that will guide you in answering these main questions and writing the report. It consists of three sections: In Section 3 you will use Weka to get insight into the data set. Guided by this insight, you will then experiment with ID3, Naive Bayes and k-Nearest Neighbour according to the instructions in Section 4. And finally, you will write down your findings in a well-organised report that satisfies very strict scientific criteria, which are described in Section 5.

3 Exploratory Data Analysis

First, open the dataset in a text editor[1] and examine it.

- Which features are included in the data set and what are their possible values?
- Which are the possible classes?

Next, start up Weka. When the GUI Chooser appears, select the Explorer. (In this assignment, we will only use the Explorer, but Weka has other functions for conducting extensive experiments under, for example, the Experimenter.) On the Preprocess tab, choose Open file and select splice-train.arff. You can analyse your dataset before classifying anything.

- How many examples are there in the train set?

If you select an attribute, then the coloured bars show you how often each class occurs for each value of the attribute. Use "Visualize All" to compare these bars for all attributes.

- Do you notice any differences between the attributes? Do any attributes stand out as being different from the others?

On the Select attributes tab, select the InfoGainAttributeEval Attribute Evaluator and Ranker Search Method. If you click Start, this will estimate the mutual information (i.e. the information gain) between each attribute and the class labels from the train set. Inspect the information gain.

[1] Do *not* use Word, because it does not show you the files as they really are! Wordpad is fine.

- Do you notice any differences between the attributes? Do any attributes stand out as being different from the others?

Finally, on the Visualize tab, you can plot various attributes against one another, or against the class label.

- What is being plotted?

Experiment with Weka until you can answer the following questions:

- Which attributes do you think will be most useful in classification? Why?
- Are there any attributes that might confuse any of the classifiers?

You must include results and/or plots from Weka to support your claims.

4 Algorithm performance

Next, go to the Classify tab. You can see a button for selecting the classifier to use (Naive Bayes is called NaiveBayesSimple and k-Nearest Neighbour is called IBk in Weka). Once you have done that, you can change the options for the classifier by clicking on its name, which is written in bold letters. Use either the validation set or the test set as your Supplied test set. Then press Start to have Weka classify the data. Weka will report many statistics, but it will be sufficient to examine the number of (in)correctly classified instances.

In the following subsections, you will first use the validation set to select an appropriate value for k and maybe remove some attributes. Then you will use the test set to compare the performance of the algorithms.

4.1 Selecting Parameters and Attributes

First select the validation set to test on. Now answer the following question:

- For the k-Nearest Neighbour algorithm, how does the value for k influence the performance of the algorithm? Support your claims by including a table with results for different values of k.

Think about whether you want to report percentages of correct classifications or absolute numbers. You should select the values of k that you try in a structured and organised way.

- Decide on a single value for k that you think is best. You will not be allowed to change it later on after you have seen results on the test set!

Now try the ID3 algorithm. You will find that it cannot classify most of the data in the test set.

- How can this be? Hint: Maybe some of the attributes are confusing the algorithm...

You can remove attributes in the Preprocess tab. If you do this, Weka will complain that the test set is no longer compatible with the train set. Hence you will also have to remove the same attributes from the test set by opening it under the Preprocess tab and saving it again after removing the attribute. (Make sure to select a file format; otherwise saving will fail without any error message.) For each of the algorithms, try removing some attributes in the Preprocess tab. You are allowed to remove different attributes (or no attributes at all) for each algorithm.

- Does removing attributes improve performance? What causes this?
- For each of the algorithms decide which attributes you will use. You will not be allowed to change them later on after you have seen results on the test set!

4.2 Final Evaluation

Now run the ID3, Naive Bayes and k-Nearest Neighbour algorithms on the test set and analyse the results:

- Create a table that reports the performance of each of the classifiers.
- Which algorithm scores best? Try to explain this.

5 Report Your Findings

Write a report with the main purpose of answering the main questions from Section 2. Your report should also answer all the other questions from this assignment, and include any other interesting things you have noticed. **Warning: Your paper cannot be over 4 Word pages or 6 LaTeX pages in length.** Any half page over this limit will reduce your grade by 2 points!

You should organise your report in a scientific way. This means that tables and figures should all have a caption (description) and a number (e.g. Figure 3), and you references to them are not allowed to depend on their position on the page. For example, you may write: "Table 2 shows the performance of the Naive Bayes algorithm on the test set.", but not: "The following table shows the performance of the Naive Bayes algorithm on the test set." Furthermore, your report should consist of the following sections, in this order:

Abstract: Explain briefly (a guideline would be approximately 75 words) your topic of research and your findings. Make sure to give away the punchline here! (You are not writing a novel, in which you want to keep your reader guessing what's going to happen next.)

Introduction: Introduce your research, explain which main questions you are going to answer and say briefly what your answers will be, discuss what each further section will cover.

Experiments and results: Discuss (in multiple sections, if necessary) your research and the experiments you have performed. What have you investigated? How have you investigated it? It is important that you include sufficient details about what you have done to allow others to reproduce your results if they want to. You should also include your results. For all experiments, make sure to clearly separate the description of each experiment from its results. Consider displaying results in tables and graphs; These are easier to comprehend.

Discussion and Conclusion: Summarise and analyse your findings and draw conclusions. It is important that your conclusions are clearly separated from your results: Results are facts, your interpretation of them will to some extent be your opinion. Come back to the research questions from your introduction. What are your answers and how are they supported by your results?

References: Finally, include a list of the references (books/papers/anything else) you have used to prepare your report. This will probably just be Mitchell's book, and maybe Tim's slides.

If you want an example of a scientific paper, you might take a look at `http://www.cs.vu.nl/%7Egusz/papers/2006-SAB-SwarmRobotics-EibenNitschkeSchut.pdf`.

You should submit your report through BlackBoard before the deadline of **December 6th**. We prefer PDF documents, and cannot accept docx (Office 2007) as we have no way of opening them.

6 Grading

You will be judged on:

- The quality of your answers to the questions (1/3 of your grade);
- The overall clarity of your report (1/3 of your grade);
- Whether your report is organised in a scientific way (1/3 of your grade); and
- You can get extra credit for doing more than necessary.

Good luck!