DME: Instructions for Mini-Project

Charles Sutton

Spring 2013

The goal of the project is to compare a number of machine learning methods on a real data set. A list of potential data sets is available on the DME web page. For each dataset, the course web page gives a description of the task to be undertaken. Typically, these are classification or clustering tasks. Each task will contain a pointer to a previous good method for this data set.

Essentially, your project should involve

1. Exploring the dataset to determine which methods and features are likely to work well
2. Reading a few reports of previous groups that have worked on this task. An extensive literature search is not expected, 2-3 papers is OK.
3. Choosing some methods that might work well on this task, based on the first two steps
4. Evaluating the results of the different methods on the task. Based on your results, are you confident which method is best?

The use of the term "method" is chosen to be deliberately vague. A typical project might compare a number of different classification algorithms on the data set. Alternatively, you might instead choose to compare different methods of feature selection or different hand-built choices of features. You might choose to generate learning curves to find out the extent to which the performance of the classifier degrades as a function of the training set. If you are extremely ambitious, you might think about whether there are advanced techniques from PMR or MLPR that you could try, or whether you can find any other information about your data set, e.g., from the Internet, that you could use as features. You are certainly NOT expected to do all of those things in one project. You are not required to use the "previous good" method for your data set; this is intended only as a source of ideas.

Essentially, you are expected to use your insight and imagination to try something that you think will perform well on the task. If your ideas doesn't work out, that is OK, as long as they were reasonable and evaluated properly. Whatever you choose to do, you must make statistically rigorous comparisons based on the principles that we have discussed in class.

You are not expected to implement your own learning methods or to develop new methods, although you are allowed to do so if you wish. You are also not expected to match the best published performance on your data set in the amount of time that you have available. However, a good project will discuss: How do your numbers compare to the numbers in the previous work? Is simply comparing the numerical results fair? (There are various reasons why it might not be, and that's OK.) What are the most important things you would do next if you had time?

All project will typically include some amount of exploratory data analysis, using graphics or summary statistics, as appropriate. This you help you decide which methods or features to use.

Finally, some of the data sets may be too large to use directly in the software that you have. (Weka is particularly bad in this regard.) In such cases, you are allowed to subsample the data set so that it will fit in memory.

I would expect the mini-project to take you around 35 hours work. You may work in groups of at most 3 people.

Choosing your project and group: **By 4pm on 11 Feb**, please send an email to the TA that says which project you would like to work on. If possible, please have your group organised by that time. List all the group members in your email, and I only need one email per group. If you do not have a group yet, **you still need to send an email** telling me which data set you want to work on, and whether you want to be in a group. If you want to be in a group and cannot find one, we will match you up. Each task should contain a pointer to a previous good method for this data set. With the consent of the instructor, you may choose a data set not on the list.

Interim report: By **4 pm on 28 Feb** you must email the instructor a 1-page ascii description of your progress so far and your plans for completion of the mini-project by the final deadline. You should discuss what comparisons

you want to run. This report will not form part of your numerical mark for the course. The goal of interim report is to make sure that your project has the right scope and that you are on track.

Final due date: Evaluation of the work on the mini project will be by a written report. This is due by manual submission to ITO by **4pm on 21 March**. Each group need only submit one report.

Late penalties: The policy of the School of Informatics is that no late submissions are allowed except on valid ground agreed a priori with the year organiser.

The report. The report should be around 6-12 pages in length of single spaced text. The following headings are likely to be useful.

- Abstract
- Overview of the task
- Previous work (literature review)
- Data preparation
- Exploratory Data Analysis
- Learning methods used
- Results, evaluation
- Conclusions

Marking Breakdown

The marking criteria include the appropriateness of the machine learning methods chosen, quality of the analysis, the quality of the evaluation, the amount of work, and the quality of the explanation of the report (both text and graphics). A guide to the letter marks are:

A Well explained description of points above plus extra achievement at understanding or analysis of results. Clear explanations, evidence of creative or deeper thought will contribute to a higher grade.
B Well explained description of points above.
C Good description of points above but significant deficiencies.
D Evidence that the student has gained some understanding, but not addressed that specified task properly.
E/F/G serious error or slack work.

Plagiarism policy

The projects are (usually) group projects. Hence you are expected to discuss the work within your group, and to work on your report together. You should write up the project as a whole, including the work of the others in your project. At the end of your report, I ask that you include a short note stating how you distributed the work amongst your team.

See `http://www.inf.ed.ac.uk/admin/ITO/DivisionalGuidelinesPlagiarism.html` for further information.