IRDS: Instructions for Mini-Project

Charles Sutton

Autumn 2014

The mini-project is an open-ended project in which you should compare a number of data science methods on a real data set or a realistic scenario. You choose which types of methods and which data sets you wish to compare. For example, you may choose to compare methods in machine learning, statistics, optimization, natural language processing, computer vision — anything that falls within the broad remit of "data science".

Essentially, your project should involve

1. Exploring the data set to determine which methods and features are likely to work well

2. Choosing some methods that might work well on this task, based on your exploration of the data set or of previous methods that have been tried on this task.

3. Evaluating the results of the different methods on the task. Based on your results, are you confident which method is best?

Because the definition of the project is deliberately vague, here is an example to give you a sense of how much effort is expected. You might choose a data set, split it into a training and test set, visualize some of the features that are most closely associated with the class label, and compare several different classification algorithms on the data set, choosing their tuning parameters carefully (e.g., by cross validation).

Your project need not follow this template. For example, you might instead choose to compare different methods of feature selection or different hand-built choices of features, rather than focusing on the choice of classification algorithm. Additionally, 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. Or you might try to find additional information from the Internet that you can use as features. If you are extremely ambitious, you might think about whether you could use more advanced ML techniques, but then you would need to be sure that you have a good existing implementation available — you will not have time to implement and debug something complex as part of this project. Note that this list is intended as a source of ideas to choose from; you are certainly NOT expected to do all of those things in one project.

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 don't work out, that is OK, as long as they are reasonable and properly evaluated.

Projects 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.

Overall I would expect the mini-project to take you around 30 hours work.

Choosing Your Project. A list of potential data sets and projects is given on the IRDS web page. You may choose one of these, or propose your own project if you like. If you wish to propose your own project, please discuss it with me, so that we can make sure it is feasible in time that you have available.

Existing Software. 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 may use any existing software that you like. In your report, you should be clear about what software you used from others, and what you did yourself as part of the project.

Comparison to Previous Work. You are 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?

Schedule

Choosing your project and group: **By 4pm on 24 October**, please send an email to the instructor and TA that says which project and data set you would like to work on. If you choose one of the suggested projects from the web site, then just give the title. If you are choosing a data set not on the list, then please write a paragraph or two that explains the data set and task, similar to the descriptions on the web site. It would be good to chat with me first before writing this description.

Interim report: By **4 pm on 7 November** 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.

Tutorials: The project will be supported by tutorial meetings on 31 October and 7 November. These will be informal groups in which you can discuss your projects with your fellow students, and share ideas and advice.

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 November**.

Poster Session: Finally, you will present your work during a poster session on the afternoon of **28 November**, to share the results more generally.

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.

Report Structure

The report should be around 6-8 pages in length of single spaced text. The report should describe what the problem was, what you did, why you made the decisions that you made, and what happenened. The exact structure of your report will depend on your specific project, but 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 project will be marked 80% based on the written report, and 20% based on a poster.

Written Report

The marking criteria include the appropriateness of the 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.

Poster Session

Presenting a poster is an important skill in research; many of the main conferences in data science feature poster sessions.

You should create an A0 size poster about your project, which will be presented as part of the poster presentation at the end of the course. Your presentation should be targetted at your fellow students; if they can understand your work, then hopefully I will too.

Your poster will be marked primarily on the quality of its presentation, that is, how effective it is on helping people to understand the main ideas of your project. Posters will receive a higher mark if they state the main ideas clearly, without excessive detail, use graphs where appropriate to communicate results, and are easy to view from a distance. It is unnecessary to spend inordinate time on the graphical design. "Simple but clear" should be your goal.

At a general level, your poster will be likely to contain the same set of sections as your report, although in each of the sections, the poster will focus on the big picture rather than all of the details. Including examples of the data is a very good idea.

The course lecturer will visit all of the posters during the session in order to aid in the assessment.

Advice on producing a good poster will be available elsewhere on the course web site.

For information about having your poster printed, please see

`http://www.inf.ed.ac.uk/student-services/joint-centre-for-doctoral-training-information/Poster%20Printing`

Please note that the KB Copy Shop will require advance notice. If you need to print a poster at the last minute — not that any of you would do this, of course — you may also consider the A0 printers in the Main Library which are very convenient.

Policy on Collaboration and Plagiarism

I encourage you to discuss your projects with other students in order to share ideas and ask questions. Indeed, we will have tutorial groups that are specifically designed for you to do so.

That said, the work and the writing that you present should be your own. For more information about the School Plagiarism policy, see `http://www.inf.ed.ac.uk/admin/ITO/DivisionalGuidelinesPlagiarism.html`.