Machine Learning Coursework Description

Simon Rogers and Maurizio Filippone

February 11, 2014

1 Introduction

The *division* is a key aspect of parliamentary democracy in the UK. In a division, Members of Parliament (MPs) vote for or against a motion (to, for example, change a law). In a typical Westminster parliament (approximately 5 years), there will be well over 1000 such divions on many different issues. For more information on divisions, see e.g. parliamentary divisions.

This coursework consists of two tasks based on data from such divisions from the UK Parliament that ran from 2005 until 2010.

2 The data

The data are available on Moodle – `courseworkData.zip`. The archive consists of four files:

- `divisions.csv` – a text file with text descriptions of what was being voted on (this file isn't necessary for the coursework, but might be interesting).
- `trainData.csv` – a csv file with one line per MP. The first column gives the MPs name, the second their party and the next 1288 their votes. A vote of 1 is a vote for the motion, -1 a vote against and 0 means that they didn't vote.
- `testData.csv` – a csv file including the test data but with the MPs parties (and names!) blanked out.
- `gapData.csv` – a csv file in the same format as `trainData.csv` but with some values missing (missing values are denoted NaN) (see task 2).

3 The Tasks

3.1 Task 1

The first task involves building a classifier to determine which party an MP belongs to based on their voting habits. This will be evaluates on a test dataset which will be in the same format as trainData.csv but will contain previously unseen MPs. Some comments:

- You are free to use any algorithm you choose, implemented in any language you choose.
- It is up to you how you handle the missing votes (the zeros) – you could ignore those (MP,vote) pairs or include them as data (perhaps there is information in the missing values).

As well as your code and report (see below), you should submit a .csv file in exactly the same format as testData.csv but with the parties that you have predicted filled in. The filename **must** be:

`predictions_<matric>.csv`

i.e. "predictions" followed by an underscore, followed by your matric number. Remember to leave a blank column at the start where the MPs name would be!

3.2 Task 2

The second task involves imputing some missing values. The file `gapData.csv` is in the same format as `trainData.csv` but some of the (MP,vote) combinations have a value of NaN. These NaN correspond to ± 1 (i.e. none of them were 0s). The task here is to predict the values that these should take. You should submit a CSV file of exactly the same format but with the gaps filled in. The filename **must** be:

```
gaps_<matric>.csv
```

4 How will it be assessed?

4.1 All students

You will submit code, csv files as described above, and a report. Your report should document how your model works, any assumptions you made, and the procedures you followed to optimise parameters etc. Your report should be pitched at a level that a mathematically literate final year student could understand.

4.2 Level M students

You need to additionally include a short (approx 1-2 pages) literature review. This could, for example, be focused on analysis of this particular dataset or analysis of data with these characteristics. Your literature review should only cover one of the two tasks (your choice which) – please make it clear which you have covered.

5 Deadline

4:30pm, 28th February 2014. Submission will be online – more details nearer the time.

5.1 Mark scheme

- (20%) Code, and correctly formatted predictive .csv files. Your predictions don't have to be particularly good, but it must be clear from the code that you have attempted something! We do, however, reserve the right to display a predictive league table in a lecture.
- (20%, Level M students only) Literature review (1 to 2 pages). Include this in your report (below) – do not submit as a separate document.
- (80% (L4), 60% (M)) Report:
  - (25% of report total): Description of problem and high level description of chosen solution.
  - (25% of report total): Discussion and justification of assumptions in model and analysis.
  - (25% of report total): Description of procedures used (e.g. models, inference techniques, analysis of results).
  - (25% of report total): Overall written report quality. Clarity of writing, use of visualisations etc.