Coursework 1: Predictive Data Mining Group Project

Brief

Effort: ~40 hours per student
Credit: 30% of overall module mark
Team size: 5 or 6 students.

Handin #1

Due date: Friday 17th February 2017, 16:00.
Handin: 1617/COMP6237/1/ (http://handin.ecs.soton.ac.uk/handin/1617/COMP6237/1/)
Required files: brief.pdf

Handin #2

Due date: Wednesday 10th May 2017, 16:00.
Handin: 1617/COMP6237/2/ (http://handin.ecs.soton.ac.uk/handin/1617/COMP6237/2/)
Required files: paper.pdf; mark_split.pdf

Overview

In this coursework you form a team and choose a predictive data mining problem to tackle. Teams are expected to perform a series of experiments on their chosen problem in order to build a predictive model (or set of models) which they evaluate and compare (across techniques used by the team and any other published approaches). The team must present their work as a written conference paper together with an associated oral presentation.

Details

Students will form groups of five or six members. Each group will propose a project in consultation with members of staff; we expect that you'll pick a dataset or challenge that has been or is being used for current data mining research (see the links at the end of this document for some ideas). The project must demonstrate the team's ability to scientifically tackle a real-world *predictive* data-mining problem (as opposed to a *descriptive* or *understanding* type problem that you'll tackle in the individual coursework).

Deliverables

There are four deliverables for this coursework:

- Project Brief. Teams must submit a short (1 side A4 max) project brief to ECS handin by Friday 17th February. In addition to describing the proposed project, the brief must include the names and email addresses of all the team members.
- Project Pitch + Q&A. In the afternoon slot on the Friday of week 4 (24th Feb) the groups will provide a 2 minute pitch of the project to the class, followed by 3 minutes of open discussion/questions.
- Conference Paper. Teams must submit a final report in the style of a conference paper by 16:00 on Wednesday 10th May to ECS Handin. The paper must use the ACM proceedings style (https://www.acm.org/publications/proceedings-template) and be at most 6 pages in length, including all references and appendices (if used). Additionally each team is required to submit a proposed marks distribution form (see below).
- Project Presentation. Teams will present their work during the lecture/tutorial slots from the end of week 11. To ensure fairness, all teams must be prepared to present in the first of these sessions, and the

presentation order will be picked randomly during the sessions.

Notes

- Students taking COMP6208 Advanced Machine Learning must ensure that they are tackling a different problem to the one they choose in that module.
- Teams can book time to meet with Markus or Jon (ideally in the unused lecture slots) should they need help or guidance. Enrique is available at the Zepler help desk every Monday and can also offer advice.
- ECS Handin wasn't really designed with group submissions in mind; each team should nominate a team leader to make the submission on behalf of their team. The other team members will see that handin says that their submission is overdue, however this can be safely ignored.

Marking and Feedback

Each team will receive an overall mark (broken down into sub-categories). Individual marks will be assigned based on a split decided by the team. Full details below:

Learning Outcomes

- Solve real-word data-mining, data-indexing and information extraction tasks
- Demonstrate knowledge and understanding of:
    - Key concepts, tools and approaches for data mining on complex unstructured data sets
    - State-of-the-art data-mining techniques
    - Theoretical concepts and the motivations behind different data-mining approaches

Mark Scheme

The conference paper and presentation will be marked as a single piece of work using the following criteria:

Criterion	Description	Marks
Experimentation and Analysis	Analyse the problem and results obtained	35
Application of techniques	Show ability to apply predictive data mining techniques and preprocessing operations	35
Reflection	Reflect on what the experimental results tell us about the problem and the techniques used	20
Reporting	Clear and professional reporting	10

Standard ECS late submission penalties apply.

Written group feedback will be given covering the above points, and will be emailed out once marking is complete.

Marks split

Team members should agree between themselves as to how the marks awarded for the team submission will be divided between the team members (see below for instructions on how to proceed if this is not possible). The Team Leader should print out the form here (marks_split.pdf) , complete it as agreed and arrange for every member of the team to sign and date it. The completed signed form must be submitted via the ECS Handin system with the conference paper. An incomplete form (e.g. with missing signatures) means that the entire ECS Handin submission is incomplete and therefore subject to penalties.

Teams are encouraged to split the work evenly between all team members (in which case the marks split evenly). They are advised to consider any proposed non-uniform distribution very carefully before submission. Note that an individual contribution of zero is acceptable and will result in that team member being effectively removed from the team. One or more individual contributions of 10% or less may result in an ad-hoc reduction in the effective team size. Any proposed non-uniform distribution will be discussed with the team after the presentation and may be subject to modification by the Module Leader at that stage.

'Failure to Agree'

Teams are advised to make every effort to agree on the marks distribution because failure to agree will be interpreted as demonstrating a general lack of competence. However, the procedure to follow if there is no agreement is set out below:

The team should divide into two or more subteams (in the worst case, a team of size 'N' could have 'N' subteams). Each team should elect a subteam leader, who should make a full submission as detailed above. Each marks distribution form submitted should indicate proposed percentages of the overall team marks to be allocated to the members of that subteam, with a written one-page explanation of why such an allocation would be appropriate. It should be noted that any attempt by a team member to exploit the advice above (that teams should make every effort to agree) by, for example, refusing to sign the marks distribution form will not be successful (in the unlikely event that this happens, each individual should make a brief signed statement as to the facts of the case and submit this with the other documentation).

The final marks breakdown for a team that fails to agree will be determined by the Module Leader, taking all relevant factors into account. This decision will be final.

Useful Links

The following list has some pointers to places where you might get some inspiration for data mining challenges together with associated data such as evaluation criteria and comparative performance data:

- https://www.kaggle.com (https://www.kaggle.com) - source of lots of different data mining competitions.
- http://www.drivendata.org (http://www.drivendata.org) - source of lots of different data mining competitions with an emphasis on saving the world.
- http://multimediaeval.org/datasets/ (http://multimediaeval.org/datasets/) - a range of data and evaluation criteria for different types of data mining problems involving multimedia and multi-modal data.
- http://www.kdnuggets.com/competitions/past-competitions.html (http://www.kdnuggets.com/competitions/past-competitions.html) - list of past data mining competitions; data and evaluation criteria is likely to be available for many of these.
- http://webscope.sandbox.yahoo.com (http://webscope.sandbox.yahoo.com) - publicly available research datasets from Yahoo!
- http://www.kdd.org/kdd-cup (http://www.kdd.org/kdd-cup) - KDD Cup is an annual data mining competition run by ACM SIG KDD; datasets, evaluation criteria, and info previous winners are available (note that the most recent competitions are actually hosted on kaggle.com).

Questions

If you have any problems/questions then email (mailto:jsh2@ecs.soton.ac.uk) or speak to Jon (http://ecs.soton.ac.uk/people/jsh2) in his office or after the lectures.