Assignment 2: Practical Workplace–Related Data Analytics Project

32513 Advanced Data Analytics Algorithms
Spring 2018

The goal of this assignment is to develop your skills in a practical data mining project related to your workplace. If you're not working or there isn't a suitable problem in your workplace, then you are free to choose a workplace you're interested in. The main thing is to make it as practical and useful as you can. If you have trouble finding a suitable dataset, speak with me and we can find one. The goal will be to derive a business problem then solve it using data analytics. You can do this assignment by yourself or in pairs. You will need to submit a report of around 20–30 pages describing the problem and your solution.

The main thing is to choose a project that you're interested in and passionate about.

I suggest to follow the CRISP–DM approach for doing data mining. You will need to identify what is the "business problem" you are solving and what is the matching "data mining problem". You will also need to describe the data and domain, any data pre-processing you applied, what models you applied and why and the results. It is not enough to just use one classification algorithm and write the output from the Rattle window. You need to explore the data, systematically try several algorithms and parameter settings to find the best (by evaluating the quality of the classifiers) and then provide a recommendation.

What to Submit

There are four parts to your deliverable.

1. a description of the business problem you are solving together with how this translates into a data mining problem (1 page);
2. a description of your exploration of the dataset highlighting interesting or important things you found (roughly 5–10 pages with figures);
3. a description of how you approached the problem, which algorithms you looked at and the parameter settings you used (10 pages);
4. your recommended classifier with reasons why (1 page).

Hint: Think of yourself as a data analyst. Your report is being written for the admin of your institution or your customer, who does not understand data analytics in any great details.

Due Date

Due date 11:59pm 5 Oct 2018.

How to submit Please email a soft copy on UTS Online. Make sure you put your student number and your name in the document.

Extensions may be granted for assignments after consultation with the Subject Coordinator before the due date.

Late assignments will have 20 percentage points deducted from the total worth to the assignment per day late or part thereof, more than five days late the assignment will receive zero. Special Consideration, for late submission, must be arranged beforehand with the Subject Coordinator.

Assessment

Group work This assignment may be done individually or in pairs. Conditions for group work are described in the subject outline. Except for exceptional circumstances (ie. where problems occur in the group), each member will receive the same mark. If there are problems in your group, please see the Subject Coordinator.

Return I will endeavour to return marked assignments within three weeks.

Contribution to final mark This assignment contributes: 30% towards your final mark.

Objectives This assignment supports objectives 1, 3 and 4 and Graduate Attributes C2 and E1 in the subject outline.

Academic Standards Please see the subject outline for details on the ethical standards we expect from you.

Hours An average student should expect to spend around 48 hours to get a 50P result on this assignment.

Marking Scheme

Your assignment will be marked based on the following rubics.

* Problem understanding (10)
- Is the business problem understood well and clearly stated in the report? (1-10)

* Data Exploration (30)
- Is there any challenge in preprocessing the data, such as loading structured data into mathematically manageable format / examination of the scale and values of attributes? (1-10, a straightforward load of csv without mentioning any checking of attributes gets 3, missing step and treats the data as if they are readily in memory gets 1)
- Is there detailed discussion on the data characteristic related to the problem, such as how the attributes / other aspects may facilitate or pose challenges? (1-10)
- Is there analysis how to deal with / make advantage of the specific data characteristics and therefore to design a solution path? (1-5)
- Clarity in presentation, general logic etc. (1-5)

* Methodology (40)
- At least two different approaches have been tested, with reasonable motivation, experiment design and criteria for the outcomes (1-10)
- Model selection, (hyper-) parameter tuning scheme for each algorithm (1-10)
- Discussion of the time and memory complexity, with considering possible scaling up of the problems (1-10)

- Sophistication of algorithms (1-5, we reward efforts in implementation)
- Clarity in presentation (1-5, consider the tables and figures for results).

* Understanding and recommendation (/20)
- How convincing discussion on the experiment outcomes is, considering both technical soundness and narrative explanation (1-10)
- Does the final recommendation fit the problem (1-10)