Applied Data Science: Coursework Brief

Overview

The goal of this coursework is to give you experience of the lifecycle of developing a data science application. The intended learning outcomes for the coursework are the same as those for the unit as a whole:

- To acquire a working knowledge of practical data science, applied to real world problems.
- To be able to start from raw data and deliver a representation that allows a better understanding of the topics in the data.
- To have experience of using software tools for data pre-processing and management.
- To acquire first-hand experience in specific techniques for data storage.
- To understand the differences between different visualization strategies to efficiently explore the data.
- To have learned how to present and interpret data to/for a non-technical audience.
- To be able to share data under privacy constraints.
- To have practiced teamwork and time management.

Your tasks in this coursework are:

- Build a data science application that extracts useful findings from a real-world data set.
- Work effectively as part of a group.
- Follow a sound development process (e.g., see the CRISP-DM standard) taking due account of privacy and ethical issues.
- Develop and demonstrate models, analysis, and visualization that will generalize well to new data.
- Deliver an informative interim presentation of your work in progress.
- Submit a clear and useful report describing your work and your findings.
- Write a thoughtful and interesting reflective discussion of the project.

This coursework has three deliverables:

1. Each *group* will submit a single short (~10 mins) work-in-progress video presentation ahead of the penultimate week of the coursework period. Feedback on this presentation will help the group improve their final report.
   - This *video presentation* is worth 10% of the overall coursework mark.
2. Each *group* will submit a single written final report at the end of the coursework period (max. 10 sides of A4 including references, submitted in IEEE conference format).
   - This *final report* is worth 90% of the overall coursework mark.
3. Each *individual* will submit a reflective discussion of their work (max. one side of A4 submitted in IEEE conference format), describing their role in the group, what their achievements were, and what they have learned from the course. The quality of this piece of work, combined with their supervisor's assessment of their contribution to the project will determine the final mark that each individual receives.
   - This *reflective discussion* can only increase an individual's mark above that achieved by the group for deliverables 1 and 2.

Video Presentation:

Your group's video presentation should be around 10 mins and should contain the following:

- <u>Introduction:</u> The background to the problem tackled by the project.
- <u>Initial Thinking:</u> What you decided to focus on and why.
- <u>Work So Far:</u> The preliminary findings that have been obtained so far.
- <u>Things To Do:</u> What remains to be done.
- <u>Challenges:</u> Obstacles encountered and how you are approaching them.

The (equally weighted) marking criteria for the video are:

- <u>Sound Rationale:</u> Decisions and direction taken are explained well and make sense.
- <u>Technical Command:</u> Treatment of technical issues is detailed, deep and convincing.
- <u>Effective Teamwork:</u> Good evidence of collaboration, planning, communication.
- <u>Good Presentation:</u> Video is succinct, clear, well-structured, comprehensive.

Note: the video itself need not have high production values or be particularly complex. An efficient, clear and informative video is what you should be aiming for, not one that is fancy.

Note: the point is to enable us to give you helpful feedback that will improve your final report. Feel free to ask for advice on specific choices that you are making or approaches that you are considering. An honest "warts-and-all" account is better than a positive but partial story.

Final Report:

Your group's *final report* should contain the following:

- <u>Abstract:</u> A succinct account of the project and its findings.
- <u>Introduction:</u> The background context for the problem tackled by the project.
- <u>Data Preparation:</u> What data manipulation was necessary to prepare a dataset.
- <u>Data Exploration:</u> What you learned from your initial exploration of the data.
- <u>Data Modelling:</u> Details of the models, visualization, and analyses that you developed, their strengths and weaknesses, and justifications for your approach and choices. You should include a technical description of any algorithms, tools, methods, methodology used (referencing the scientific literature or equivalent wherever appropriate).
- <u>Findings:</u> Details of your findings, results, insights, etc. Use appropriate visualizations.
- <u>Discussion and Conclusion:</u> What has the project delivered in terms of knowledge, understanding and insight into the problem that it addressed?

The marking criteria for the final report are:

- <u>Problem Understanding:</u> To what extent have you understood and explained the goals of the project, taking account of end-user requirements (15 marks).
- <u>Data Wrangling and Exploration:</u> To what extent has extraction, preparation and manipulation of the data been robust, systematic and sophisticated; Depth of insights gained through the initial exploration of the data (15 marks).
- <u>Data Modelling, Visualization, Analysis and Evaluation:</u> Effectiveness of tools chosen and deployed; Appropriateness of the modelling approach; Soundness of statistical and machine learning methods applied; Sophistication of the visualizations employed; Effectiveness and insight of the evaluation (40 marks).
- <u>Discussion and Conclusion:</u> Utility of the project findings for the end user (15 marks).
- <u>Presentation:</u> Fluency and coherence of the written text; Quality of the figures, charts, graphs, etc. (15 marks).

Reflective Discussion:

Each group's report will be marked using the criteria described above and this will be used to set an overall mark for each group. Each individual's mark will be determined based on this overall mark (so it is in everyone's interests to collaborate well) with a potential bonus based on the individual's contribution. This will be assessed partly by your attendance at and contribution to meetings with your group's supervisor and partly based on a 1-page reflective discussion document that describes your role in the group, what your achievements were, and what you have learned from the course.

Your reflective discussion could contain the following kinds of material, but the content is up to you:

- Account of your personal contribution to the project.
- Discussion of the overall strengths and weaknesses of the project.
- Reflection on how the project related to the ideas taught on the unit.
- Discussion of your achievements and what the project experience taught you.

The marking criteria for the reflective discussion are:

- Significance of contribution to the project.
- Appreciation of the project's strengths and weaknesses.
- Appreciation of how the material taught on the unit relates to the project.
- Reflection on achievements and what was learned from the project experience.

Marking for COMSM0055 and COMSM0056

The way in which the marking criteria for all three deliverables are applied for students in the fourth year of an undergraduate degree or in the first year of a CDT programme will be consistent with the elevated expectations associated with an M-level unit.

In such cases, work will be expected to exhibit a deep and mature level of analysis and reflection, and the application of tools and methods will be expected to be accomplished and their evaluation and interpretation sophisticated.

For example, for a solid piece of M-level work:

- Machine learning methods will be expected to be used in a way that demonstrates command of the methods and the ability to tailor them to the problem at hand, rather than merely exploit off the shelf techniques.
- Visualization should be carefully considered in terms of accurate and unbiased representation of the data and some innovation in terms of visualization methods may be expected.
- Analysis of findings will be expected to go beyond summary statistics and qualitative comparisons to include relevant quantitative inferential statistical methods.
- Etc.