CEE 690.03: Health and Environmental Data Science

Spring 2018 Course Project

Guidance for 50% Status Report:

Due date: 11:59PM Tuesday, April 11th
Submit a file electronically on Sakai. All standard formats are okay.

Brief Description

The goal of this status report is to get of the material for the final report together and make sure that each individual or group is on track for the final report. This involves talking about the progress so far, and the plan between now and the final report.

Note that it is perfectly permissive to take your own writing from the project proposal to use in this document. Furthermore, much of this document can be reused for your final report, so take care to set this up so you can reuse your work.

If there was an issue raised in the feedback of the project proposal, this must be addressed. If it is not addressed, you will lose *additional* points in that category (up to the max per category).

Page limit:

There is a 4 page limit to the progress report (12 point font, 1 inch margins), excluding references (references can go on additional pages, i.e. page 5 will be only references). This is essentially 2 additional pages beyond the project proposal, which should largely be able to be included in this progress report. Be *succinct*, but convey all necessary information. Conveying information effectively in a short space requires practice, so please take the opportunity to practice this important skill.

Planned Feedback:

In addition to feedback on the required components below, the major feedback will be targeted at two points:

1. Is there a feasible and thought-out plan between now and the final project report?
2. Is the plan up to the level of work expected for the final project? (Note that this was largely checked at the project proposal stage. This is expected to be a yes for every group and individual.)

One again, a reminder that this project will be evaluated on the quality and plan of approach, not on the outcome metrics of the approach. *Often we get negative results in the real world!* If your results are negative, that is fine, but you should describe why that is and what assumptions you made about the data turned out to be incorrect (i.e. why are the results negative?).

Required Components:

15%: Communication Clarity

This expectation will not change between this report and the final report.

This progress report should be clearly written, easy to follow, and complete as a stand-alone document. Think of this as delivering a report to a customer who hired you to complete the project. Alternatively, someone who has not taken this or a similar course, but it familiar with probability/statistics and linear algebra, should be able to read your status report and *conceptually* understand what you have done. Beyond this, enough detail on the actual methods should be included that an expert could easily reproduce what you have done.

This should be a formal report. Writing, including complete sentences, grammar, etc., matters. Some important reminders:

1. Don't list the questions that need to be answered in the report, and then respond to them
2. Write a stand-alone document
3. Define acronyms the first time they are used (e.g. Support Vector Machine (SVM))
4. Divide the report into sections and subsections to help improve the document flow
5. **Use a clear and descriptive title**

10%: Problem Description

This expectation will not change between this report and the final report. This section should be finalized here, and you should simply be able to copy this over to your final report.

You should include a description of the project here, and include specifically the question that you are trying to ask of the data. The document should be self-sufficient, such that someone that doesn't have a background in that problem domain can read and *understand what task you are trying to accomplish and why*.

Provide background and context for your problem. This should be more significant than "this is a well-used prediction task from Kaggle" or something of the sort. A well-written project description will typically take at least most of a page, but there is no formal limit either high or low on what's in here. Some example questions that this section should answer are:

- Why is this an interesting or important problem?
- Why is this problem significant?
- Why would this problem be interesting or significant to someone else?
- What would the potential impact of your results be?
- What is your hypothesis on the data?
- What is the relevant literature on this topic?

10 % Data Description

This expectation will not change between this report and the final report. This section should be finalized here, and you should simply be able to copy this over to your final report.

The data description should be complete and thorough. Someone reading your report should be able to achieve a good understanding of the data you are using from the written description in the report.

This includes details on what is the source of your data (including references if necessary), how it was collected or generated, and how this data relates to question you are trying to ask. This should also include summary statistics (e.g. number of data points, what the outcome/target distribution is, etc.). You should utilize visualizations to help describe the data.

10%: Preprocessing

This expectation will not change between this report and the final report. This preprocessing may still be changing, but you should write this so you can simply update it later.

The preprocessing description should include details on how the raw data was processed prior to the use of the core statistical or machine learning methodology. This can include steps such as outlier removal, data normalization, feature selection, feature extraction, etc. This should address what preprocessing steps were chosen, why those preprocessing steps were chosen, how they were implemented mathematically, and what empirical benefits the preprocessing steps convey.

Some questions this should address:

- What preprocessing steps were tried?
- Why were these things helpful or necessary?
- What is the empirical support for these steps?

10 %: Core Methods

*This expectation will change between this report and the final report, because not all details need to be filled out at this point. The **basic** strategy should be in place, so you should simply be able to mostly copy this over to your final report. I want to see that at least a simple approach has run on the data.*

Include a complete description of what statistical or machine learning methodology have you chosen for this project. This description is expected to be complete and thorough, and include a mathematical description of what the methodology is doing. Visualizations are often helpful to describe what the methodology is doing, but depending on the method this may or may not be appropriate.

Beyond a description of the chosen methodology, illustrate why you have chosen this methodology, including a description of alternatives and discussing trade-offs evaluated when making this decision.

Some key questions that should be answered:

- What type of problem are you addressing? E.g. classification, regression, forecasting, recommendation system…
- What characteristics were key when choosing the approach?
- How does the chosen approach address these characteristics?
- How did you choose model parameters?
- How did you evaluate the performance and perform model selection? (e.g. cross-validation)

10 %: Results

In this status report, you should show ***preliminary results***. These results do not have to be finalized, but should show progress towards building and evaluation an approach.

You should provide quantitative descriptions of how well the different system components perform (i.e. preprocessing, feature extraction, etc.). In addition to the quantitative description, you need to ***explain the result***. Do not simply present a series of figures, but talk about the interpretation and conclusions of the figures.

Some questions:

- Overall, how is the methodology so far doing?
- Where does the approach look promising?
- Where does the approach fall short?

20 %: Current Conclusions, Unexpected Challenges, Next Steps

Talk about your current results for the pipeline you have built so far, including strengths and weaknesses. Talk about current pitfalls, and the plan for next steps moving into the final project.

Some questions that should be answered"

- Where is the system satisfactory?
- What was an unexpected pitfall?
- What is the biggest area that needs improvement?
- How will your proposed work address these issues?
- What is the plan between now and the final report? How will work be divided amongst different group members (if a group of 3)?

5 %: Figures

These expectations will be the same in the final project.

Figures should have captions that describe the content, and each figure should be referenced and described in the text.

A basic checklist:

- Don't "print screen," but export the figures to a graphics file
- Label all axes
- Include a legend when appropriate
- Make sure that the text on the figure is large enough in the report

5 %: References

These expectations will be the same in the final project.

Include all reference when necessary to archival material. This means published books and scientific articles, not Wikipedia. Do not cite our course notes. Succinctly, include a reference or citation for

- Any statement or idea that is not common knowledge
- Any statement or idea that originated with someone else
- Any algorithmic approach you did not develop.

Any commonly used reference format is okay, but the style should be consistent throughout.

5 %: Affirm that you adhered to the honor code in the report.

Should be self-explanatory.