Project delivery minor Data Science

Project Presentations

Please prepare a Powerpoint presentation which summarizes your group's project and outcomes so far. Use the following format. You should end up with 6-10 slides + title.

1. Please identify your team (team name and member names) on the title slide.

2. (1-2 slides) Problem statement: What problem you are trying to solve. Should include quality metrics you use to measure performance/accuracy. Should **not** describe the algorithms or methods you're using to solve the problem.

3. (2 slides) Methods you explored or plan to explore. Will typically include data preparation/featurization (feature selection, feature extraction, etc.), and then the learning algorithms you tried, and possibly visualization or interaction to explore the results.

4. (1-2 slides) The tools you used, and a rationale for their use. Can cover data preparation, learning, visualization, performance measurement, etc.

5. (1-2 slides) Results (may be preliminary). Any results you have to report so far. May also be a report of unexpected challenges.

6. (1-2 slides) Lessons learned and/or plans to mitigate challenges.

You should have an extra slide comparing your results with a baseline model. The baseline model can be e.g. Logistic regression or Naive Bayes for a classification task, K-means for clustering, etc. You should be able to show a significant gain over the baseline.

Mechanics

- Plan to present for 10 minutes. There should be some time for questions in the change-over between groups.

- When you're done presenting, disconnect your machine so the next group can present, but stay at the front of the room to answer questions.

- When presenting make sure you have your presentation on a (charged) laptop and also on a backup machine.

- Check that your video output works, and that you know how to use it.

- Make sure you know how to manage multiple screens (or to disable that feature) since most laptops map the projector to a different screen by default.

- Check that your presentation renders correctly on the target machine. Editing the presentation on windows for presentation on a Mac usually breaks something. Pdf will generally render the same on both platforms.

- Someone always forgets to do one of the above and misses their presentation slot. Make sure its not your group.

- Attendance is compulsory at presentation sessions.

Research Poster

Please prepare a Poster with similar content to your presentation with these items. You can use powerpoint as per the example posters.

1. Your team (team name and member names).

2. Problem statement: What problem you are trying to solve. Should include quality metrics you use to measure performance/accuracy. Should *not* describe the algorithm or method you're using to solve the problem.

3. Methods you explored or plan to explore. May include some data preparation/featurization, and then the learning algorithms you tried, and possibly visualization or interaction methods.

4. Results (may be preliminary). Any results you have to report so far. May also be a report of unexpected challenges.

5. The tools you used, and a rationale for their use. Can cover data preparation, learning, visualization, performance measurement etc.

6. Lessons learned and/or plans to mitigate challenges.

Posters can be authored in powerpoint. It’s fine to submit a powerpoint file online, but please also submit a pdf in case there are formatting problems. 30x40" size is ideal (we will rescale if needed).

Grading

Problem Statement and Background (10 points)

Give a clear and complete statement of the problem. Don't describe methods or tools yet. Where does the data come from, what are its characteristics? (4 points)

Include informal success measures (e.g. accuracy on cross-validated data, without specifying ROC or precision/recall etc) that you planned to use. (4 points)

Include background material as appropriate: who cares about this problem, what impact it has, what implications better solutions might have? Included a brief summary of any related work you know about. (2 points)

Methods (20 points)

Describe the methods you explored (usually algorithms, or data cleaning or wrangling approaches). Justify your methods in terms of the problem statement.

It will be easiest to describe methods along your data pipeline. i.e. start with data collection, then cleaning and repair, then transformation, then analysis, and any visualizations you made. (10 points)

Parameter choices can have a big effect on performance. Make sure you know what parameters (especially defaults) that you used and that they worked reasonably well on your data. (5 points)

If your results seem inconsistent with prior work or other groups', try to figure out why, and explain in your report. But don't give a manufactured explanation. There are many plausible-sounding explanations of a data anomaly, almost all of which are wrong. Make sure the evidence really supports your explanation and not others.

Be sure to include every method you tried, even if it didn't "work" or perform as well as your final approach. When describing methods that didn't work, make clear how they failed and any evaluation metrics you used to decide so. (5 points)

Results (20 points)

Give a detailed summary of the results of your work. Here is where you specify the exact performance measures you used. Be sure to justify your measure(s) in terms of the goals of your project. Usually there will be some kind of accuracy or quality measure. There may also be a performance (runtime or throughput) measure. (10 points)

Ideally you should give results across some variations of your solution like different model types or different parameter choices. (5 points)

Please use visualizations and graphs whenever possible. Include links to interactive visualizations if you built them. (5 points)

It would be reasonable to submit your report as a notebook (Jupyter), but please make sure that you include any required files. You can also submit a separated notebook as an appendix to your report if that makes the visualization/interaction task easier.

Tools (10 points)

Describe the tools that you used and the reasons for their choice. Justify them in terms of the problem itself and the methods you want to use. (5 points)

Tools will probably include machine learning, and possibly data wrangling and visualization. Please discuss all of them.

How did you employ them? What features worked well and what didn't? (5 points)

Describe any tools that you tried and ended up not using. What was the problem?

Lessons Learned (30 points)

In this section give a high-level summary of your results. If the reader only reads one section of the report, this one should be it, and it should be self-contained. You can refer back to the "Results" section for elaborations. This section should be less than a page. In particular emphasize any results that were surprising, and if so, what your exploration of them yielded. (10 points)

You should evaluate a primary model and in addition a "baseline" model. The baseline is typically the simplest model that's applicable to that data problem, e.g. Naive Bayes for classification, or K-means on raw feature data for clustering. (10 points)

If there isn't a plausible automatic baseline model, you can e.g. compare with human performance by having someone hand-solve your problem on a small subset of data. You won’t expect to achieve this level of performance, but it establishes a scale by which to measure your project's performance. Try to use labor efficiently, i.e. if the data is mostly negative instances, use your system to predict labels and then give the human a more balanced (or more likely to be balanced based on your model) selection of instances.

Compare the performance of your baseline model and primary model and discuss/explain the differences. (10 points)

Team Contributions

Please give a percentage breakdown of the effort from each team member, and what they worked on (use personal log as base). Please discuss this within your team to make sure every member agrees with the breakdown.

NOTE: The individual contribution can have a positive or negative effect on the final grade.