

# Final Project Guideline

CS573 (Data Mining), Spring 2019, Purdue University

## Project Goal

The goal of the final project is to provide students with an opportunity of getting hands-on experience in data mining, practicing the techniques and algorithms they learn in this course in real-life data mining scenarios, and even to the extent possible to work on an open research problem. Students should work on the final project in teams of 2-4 people, and the number of participants in a project will be considered in the evaluation of the project.

## Project Topics

Students are required to submit a proposal specifying the topic that they plan to work on during the final project. Students are free to pick any topic of interest in the general field related to data mining. Students are also encouraged to connect the final project with their own research.

Here are a few examples of final projects that students can consider working on:

- **(Application of Existing Data Mining Algorithms)** Choose one interesting dataset and apply at least three different data mining algorithms (not limited to the ones that are covered in this course). Make thorough comparisons among different algorithms in terms of their formulations and assumptions, parameter tuning procedures, and the performance. Reflect on the new domain knowledge that you obtain after applying these data mining algorithms (e.g., do you get any insight about the data?). You may use datasets from Kaggle competitions. In the case where more than one team choose the same Kaggle dataset, the algorithm performance difference across different teams will be considered in grading the final project.
- **(Developing New Data Mining Algorithms for Specific Problems)** Identify an unconventional domain where data mining has not been widely applied yet (but can be), or a challenging scenario where existing data mining algorithms fall short. Design data mining algorithms for the specific domain/scenario that you have identified. If you work on applying data mining algorithms to an unconventional domain to solve a specific problem, show how the performance of your algorithms compare with the start-of-art methods (which has limited utilization of data mining techniques) for solving that problem. If you work on designing data mining algorithms for a particularly challenging scenario, demonstrate how the performance of your algorithms (can be multi-dimensional, including accuracy, computational efficiency, understandability, etc.) compare with existing data mining algorithms on that scenario.
- **(Connecting to Dissertation Research)** Consider how the algorithms and techniques you learn in this course can be used in your dissertation research. Describe a particular problem in your research where data mining can be useful. Review existing methods for that problem (if any) and state their advantages and limitations. Introduce and implement your data mining solutions for that problem, and summarize the results.

## **Project Timeline**

- March 17, 2019: Final project proposal due at 11:59pm. (Send it via Blackboard)
- March 26, 2019: Final project pitch (in class); slides are due March 24, 2019 (11:59pm via Blackboard)
- April 23 & 25, 2019: Final project presentation (in class); slides are due April 21, 2019 (11:59pm via Blackboard)
- April 28, 2019: Final project report due at 11:59pm. (Send it via Blackboard)

Note that all project-related deadlines are HARD deadline. You can not apply extension days on these due dates.

## **Project Documents**

Students are asked to submit two documents for their final project: the project proposal, and the final report. For both documents, please format your submission based on ACM SIG Proceedings Templates: <https://www.acm.org/publications/proceedings-template-16dec2016>.

### Project Proposal (due 11:59pm, March 17, 2019)

A 2-page maximum document describing:

- The composition of your team
- The topic you plan to work on for the final project; in particular, please explain why you believe the topic you've chosen is an interesting and innovative topic.
- Your plan of activities to conduct in your project (e.g., literature survey, data collection and exploration, algorithm design and implementation, evaluation, etc.)
- Your plan to evaluate the outcome of your project (e.g., what do you expect to achieve through your project? How will you measure whether your project achieve the intended goals?)
- Your project timeline (e.g., how much time will you spend on each of the activities you plan to conduct for your project?)

### Final Report (due 11:59pm, April 28, 2019)

The final report is a 10-page maximum document summarizing:

- The topic of your final project (e.g., what's the background of this study? What is the type of problems your project solves?)
- Describe your dataset
- The method you use in your project: If you use existing algorithms, which ones you use and how do you tune them? If you develop new algorithms, what are the assumptions and formulations of your algorithm, and what about implementation details?
- The performance of your method: What's the performance of your method? If you implement more than one method, compare different methods' performance. If you propose a new method, compare its performance to that of the previous state-of-the-art method.

- Discuss any new insights you obtain through this final project. This could be insight for the particular domain that you work on, or insight for how to design/train data mining models and utilize data mining algorithms in general.
- Evaluation of the outcome of your project. Use the evaluation plan you have in your proposal to guide this process, and answer the following question: do you achieve your goals for the final project as you set in the proposal?
- Contribution of each team member in the final project.

Consider to use the final report as an opportunity to practice your *scientific writing skills*. Thus, throughout the process of writing the final project report, please think about:

- How can I convince the readers that the topic of my project is an important and interesting one?
- How can I communicate the methods I use in the project to the readers in an organized and accessible way?
- How to convince the readers the findings of my project?
- How to ensure that readers of my report understand the key take-away messages of my project?

## **Project Presentations**

Students will give two short presentations about their project: final project pitch and final presentation. Note that for distance students, if you are not in a team with on-campus students, please record your presentations and submit it via Blackboard.

### Final project pitch (March 26, 2019)

Each project team gets 2 minutes to present to the class: (1) the topic proposed to work on for the final project, (2) why they are excited about this project, and (3) what's the expected outcome of this project. Each team then gets 1 minute to answer questions from the audience.

Note the specific amount of presentation time allotted to each team is subject to change. We will finalize the time limits after we get all proposals and have a more accurate estimate of the number of teams.

### Final project presentation (April 23 & 25, 2019)

Each team gets 6 minutes to present to the class: (1) what is the problem they are solving in the final project? (2) what are the methods that they take and what are their results/findings? (3) what are the insights they obtain through the final project? and (4) what are the possible future work for the project? Each team then gets 2 minutes to answer questions from the audience.

Note the specific amount of presentation time allotted to each team is subject to change. We will finalize the time limits after we get all proposals and have a more accurate estimate of the number of teams.

## **Project Grading**

Final project contributes to 25% of the final grade in this class. A detailed breakdown is as follows:

- Project proposal + pitch: 5%
- Project final presentation: 10%
- Project final report: 10%