Algorithms for Big Data – Project Information

- Instructor: Cyrus Rashtchian (personal website)
- Fall 2020
- MWF 11a – 11:50a
- Course Website: http://madscience.ucsd.edu/bigdata.html

Project Overview: The project can either be an in-depth exploration of a paper or a new research-level investigation. Overall, the final project does not need to be original to the level of a research publication, but the effort of the students should be clear. On the other hand, students are encouraged to work on projects that may turn into future publications at research venues. For theory projects, the goal is to explore the possibilities and limitations of known techniques (ideally improving known results and/or expanding knowledge to new regimes). For empirical projects, the goal is to study the efficacy and trade-offs of algorithms in specific settings (either on standard datasets or on compelling synthetic data distributions). The ideal outcome of a project consists of new insights and/or examples that shed light on previously studied problems. The following types of projects are encouraged:

- **Reading-based:** read a few recent research papers on a concrete topic and summarize them.
- **Implementation-based:** implement some of the algorithms from the class (or from other theoretical literature), and perhaps apply to your area of interest/expertise, using real-world datasets. One aspect of such projects must be comparison among a few algorithms.
- **Research-based:** investigate a research topic on your own (e.g., develop an algorithm, and prove its properties; or prove an impossibility result). It may be more applied: e.g., perhaps in your area, certain theoretical algorithms can be modified to have even better performance, due to special properties of the datasets, etc.

The topic of your project must be within the scope of Theoretical Computer Science, Data Science, or Machine Learning (please talk to me otherwise). It is expected that there will be an algorithmic component. In particular, the focus is on algorithms with provable guarantees (for the implementation type, you may compare such theoretical guarantees with heuristics though).

Group work: For the project, you may work in groups of size 1-3. It is ideal to work in groups of size 2-3.

Project Scope: The specifics of the project will be flexible. Each student should explore a few related topics, with the goal of demonstrating or extending known techniques. The tools that you explore can be either from this class or from 1-2 research papers. Regardless of the scope, the students are expected to be thorough in their exploration. This is an opportunity to dig deeply into a new area, learning about mathematical methods in the context of a concrete area.

Advice and Observations:

- Regardless of whether the project is more empirical or theoretical, you must provide a formal, mathematical model of the problem that you are trying to solve. This may contain a probabilistic model of the data (e.g., independence assumptions, distribution of data) or the system used to solve the problem (e.g., sequential, distributed, streaming). Part of the challenge of research is providing a clear and concise problem formulation that enables future work to build on the topic by using alternate approaches. This allows the results to be placed formally in the context.

- You should provide a short review of the relevant literature and techniques. This should include a discussion of why certain approaches are likely to succeed or fail for what you are trying to do. In the final report, it would be good to include at least 10 citations to relevant papers in the area (although you are only expected to read 1-2 in full detail).
- When motivating the project, you should explain how it fits into a broader picture of the research world. For example: What are some other applications that could benefit from similar techniques? How would your solution change if some of the assumptions were different? How does your problem relate to other well-studied research areas?
- For a theoretical project, it can be helpful to formulate concrete conjectures, even if you are not able to prove them yourself. Then, you can try to verify some special cases, or discuss what would happen if the conjecture were true or false.
- The project can be related to or part of a larger ongoing research project that you are working on. In this case, you should be clear about (1) what has already been done in the larger project, (2) how this fits into the larger picture, and (3) what is new and different about the class project compared to existing or previous work in the area (by you and by others).
- If you plan to compare a certain problem using multiple paradigms or approaches, then should discuss the advantages and disadvantages that you expect from each approach. This is a form of the scientific method. You may discover you expectations are wrong, and it is good to develop ideas ahead of time so you can explain why the predictions do or do not match what you find when you actually perform the theoretical or empirical analysis.
- There is a chance for the project to turn into a publication in a machine learning, data mining, or theoretical computer science venue. Realistically, this will take about 4-8 months of work after the course ends. You will have to read much more related literature. You will also have to be thorough in evaluating other approaches or baseline methods for your problem. Hopefully the mathematical formulation and preliminary results will give you a good idea about how to extend your project to be more thorough and rigorous.

Proposal (due 10/30)

In the proposal, please include the following information in a 1-2 page document (ideally in Latex)

1. What paper(s) do you intend to read? What is the broader context that these papers fit into?
2. Discuss the objectives of the project and any relevant milestones that you can complete week-by-week. There is limited time to complete the project, so please try to think clearly about concrete goals that can be achieved in the timeframe.
3. If the project is theoretical, what results do you hope to prove? What tools may be useful for proving these results? You may also want to include 1-2 previous results that you plan to improve upon, with proper citations to the relevant papers.
4. If the project has an empirical component, describe how you will evaluate your proposed solutions. What datasets will you use? If you want to use synthetic data, then what is important about the synthetic distribution? What metrics will you evaluate?
5. All research projects are required to have a clear problem statement, which is a precise formulation of the problem you are trying to solve. It is good to start thinking about this sooner than later. You should include a sketch of this in the proposal, being as formal as possible.

Progress Report (due 11/20)

This will be a 3 - 4 page document, that should be a rough draft of your final report. After working on the project for about 3 weeks, please discuss in detail what you have done so far and what is still left to do. At this point, you should have clear descriptions of the papers you are reading, as well as the problem you are trying to solve. You can also include more details and updates for the theoretical results (e.g., smaller results that you have finished or updated conjectures) or empirical metrics (preliminary experiments, baseline methods). Compared to your project proposal, this document should be much more formal and detailed, and you should take care to include relevant references and background material, e.g., you can essentially write the preliminaries section of your final report, in addition to parts of the introduction and parts of the theoretical/empirical contribution sections.

Presentation (on 12/7 or 12/9)

Each person or group will give a 10 minute presentation on their project to share with the class. This should be a "teaching" opportunity, where you can share your findings with the rest of the class. Since the time is limited, you can present the problem you are solving, briefly discuss the relevant literature, and mention 1-2 new ideas or results that you have developed. You should focus on the most important and interesting aspects, and you can include the full details in the project report.

Report (due 12/11)

This document should clearly describe the main findings for your project, with a target length of 6 – 10 pages, typeset in Latex. You can use your progress report as a starting point, and then you can add more details and further references. The report should be modeled after a typical conference paper, but it can be shorter and less polished than a published paper.

As a rough outline you can aim to have

- 1 – 2 paragraph abstract summarizing the project
- 1 – 2 page introduction, which describes the background, motivation, and novel aspects
- 1 – 2 page preliminary section, which contains the relevant mathematical details and the formal problem statement
- 3 – 6 pages of your research findings. Depending on your project, this may include descriptions of new algorithms, highly related work, theoretical or empirical results. For any theoretical results, please include full proofs of the relevant results. If you have an experimental section, then you should clearly describe the methods and datasets used, as well as provide plots and tables. All figures must have clear and descriptive captions.
- 0.25 – 0.5 page conclusion, summarizing your contributions and mentioning any immediate avenues for future work

Of course, each project is different, and you may want to follow a slightly different structure for your paper. The grading will be based more on the content than on the organization, but it is good to present your findings as clearly as possible.