COMP 441/552: Large Scale Machine Learning

Rice University

Anshumali Shrivastava

anshumali At rice.edu

09th Janauary 2017

About

- Instructor : Anshumali Shrivastava
- Email : anshumali AT rice.edu

- Class Timing: Monday/Wednesday/Friday 11am to 11:50am
- TA: Chen Luo cl67 @ rice.
- Class Location : DCH 1062
- Office Hours : TBD
- Website: `http://www.cs.rice.edu/~as143/COMP441_Spring17/index.html`
- Discussions and Announcements: Canvas

Grading Total (105%)

- Project 50% (Group of 2)[1]
- 4-5 assignments (Best 4 will be considered) 25% [2]
- 2 Quizzes 15% [3]
- 1 Scribes (Individual) 10%
- Participation and Discussion Forums 5%

[1] Grads Have Higher Bar

[2] Extra Sections for Grads

[3] Grads Will Have More Questions

IMPORTANT

- Work in Group of 2. **Both Get Same Marks, Choose Wisely.**
- Proposals Due: 23rd Jan.
- Midterm Presentation: 6th and 8th March
- Final Presentation: 19th and 21st April
- Final Reports Due: 1st May

What Should A Project Be Like?

Ideally publishable in Top Tier Conferences ICML, NIPS, KDD, etc.

- Take a popular ML algorithm with recent benchmark method/implementation. Make is (5x+) faster using parallelism/approximations. Or Reduce memory footprint.
- An end-to-end implementation of an ML algorithm with support multi-core/GPUs/Multi-node with 2-5 benchmarks. (**Less Risky**)
- Novel Estimators/Algorithms with some theoretical Analysis or Large Scale Evaluations. Must show advantage over existing methods.
- Creating (or having access to a unique) Large-Scale dataset (from mostly web), for a novel task. Organize it: label generating/creation, cleaning, etc. Run 3-4 (or more) intuitive benchmarks on it.
- **Important:** Benchmarking your proposal against 2-3 recent popular methods on performance and accuracy. Evaluations on Multiple and Large Datasets.
- Beating the best published accuracy on a popular dataset. (**Risky**)
- You can use your existing project, if it involves large-scale ML.

Projects

What should not be aimed

- Standard ML problem on existing data.
- I proposed XYZ algorithm, it works on this (small) dataset. However, there are no baselines.
- I got 5% or less improvements over standard methods on some small dataset (usually less than million examples).

Most Important Component of Class, Start Now!

How will it work

- Form a Group. I can help co-ordinate.
- Formulate the Problem and Project. Get approved by the Instructor.
- We have few pre-defined and concrete projects. Come talk.

Other Requirements

Assignments

- 4-5 bi-weekly assignments. Due on Friday in Class.
- Only 4 will be counted.

Scribes

- Each student will scribe 1 lecture, starting next week 16th.
- Scribes are due, by email, on the 5 days of the class (16th Due on 21st).
- Choose dates soon. (Spreadsheet Link Soon)
- LaTeX template on Website.

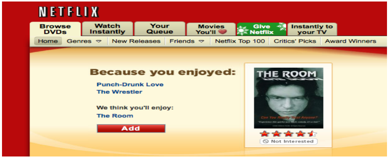
Exams

- Two 10-15 min In-Class Quizzes (**Will be Announced**).
- No Finals.
- No mid-terms.

Some Problems

- How can we search through billions of webpages quickly ?

- What goes behind recommendations engines ?

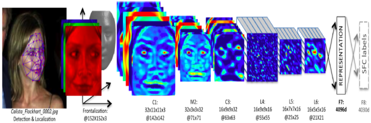


Because you enjoyed:

Punch-Drunk Love
The Wrestler

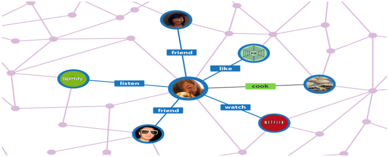
We think you'll enjoy:
The Room

- Deep Learning.



Some Problems Contd.

- How to deal with massive graphs ?



- Many more...

Some Broad Topics.

- Sketching and Streaming.
- Hashing and Randomized Algorithms.
- Optimization for Big-Data.
- Kernels Features.
- Submodular Optimization.
- Recommender Systems.
- Mining Massive Graphs.
- Deep Learning.
- Active Learning and Crowd Sourcing.
- Online Learning and Multi Arm Bandits.

Textbook

No standard Textbook: ML is a fast evolving field. Most topics are still under development.

Lecture Scribes, with references, will be made available.

You may look at

- Mining Massive Datasets (online book free)
- Scaling up Machine Learning: Parallel and Distributed Approaches (Ron Bekkerman et. al.)

Next : Some Probability