Machine Learning with Large Datasets

Logistics

Barnabás Póczos

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Logistics

Time and Location

- **Location**: Porter Hall 100
- **Time**: Tuesdays, Thursdays: 1:30-2:50pm

- **Recitations**:
  - Porter Hall 100 5:30-7:00 pm, Wednesdays.
  - 1st recitation: TBA

Class Website

https://sites.google.com/site/scalableml2018f/

Class Website

Class Announcements

Lectures

Homework & Exams

Recitations

Projects

Deadlines

Syllabus

Description

Course Staff

Office Hours

Grading Policy

Collaboration Policy

Late Day Policy

Feedback

Other Tools

We will also use

- **Piazza**: for student discussions only about Lectures, HW assignments, Projects, etc.

  TAs and the Instructor may NOT participate in these discussions.

- **Autolab**: for homework and project submissions
- **Gradescope:** for grading
- **Overleaf: for scribing** (https://www.overleaf.com/)

Please make sure you have access to Piazza and Autolab.
If you have difficulties accessing these websites, please contact the TAs.

Waitlist

- Usually many students drop the class in the first few weeks. Those places will be filled in from the waitlist.

- If you want to skip the waitlist
  - MS students: Email Dorothy Holland-Minkley, dfh@andrew.cmu.edu
  - PhD students: Email Diane Stidle, stidle@andrew.cmu.edu

  And explain the reason!

  Fire regulations: We can't have more students than the number of seats in the class room.

Auditing

To satisfy the auditing requirement, you must

- ❑ Do the midterm, final and pass.
- ❑No need to do HW or Projects
- ❑ Please send the instructors and TAs an email saying that you will be auditing the class.

Prerequisites

- **Basic ML** (E.g. 10601,10701. They can be corequisites)
- **Probabilities**
  - Distributions, densities, multi-dimensional Gaussian,…
- **Basic statistics**
  - Moments, typical distributions, regression…
- **Basic calculus:**
  - Functions, integral, gradient, convexity
- **Basic algebra:**
  - SVD, eigenvectors, orthonormal matrices, …
- **Algorithms:**
  - Dynamic programming, data structures, complexity O() …
- **Programming:**
  - Mostly Python. Matlab and Java can be also useful
- **Math:** Ability to deal with "abstract mathematical concepts"

Recitations

- **Strongly recommended**
  - Brush up pre-requisites
  - Review material (difficult topics, clear misunderstandings, extra new topics)
  - Ask questions
  - Discuss HW and Midterm solutions

- **Recitation time and location**: Friday, TBA
- **First recitation:** TBA

Textbooks

- **No required book**
- **Required reading assignments on class homepage!**

- Recommended Textbooks:
  - Scaling up Machine Learning: Parallel and Distributed Approaches; Bekkerman et al
  - Large Scale Machine Learning with Python; Sjardin et al
  - Pattern Recognition and Machine Learning; Chris Bishop
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - Machine Learning; Tom Mitchell
  - Information Theory, Inference, and Learning Algorithms; David MacKay
  - Machine Learning: A probabilistic perspective; Kevin Murphy.
  - Understanding Machine Learning: From Theory to Algorithms; Shai Shalev-Shwartz and Shai Ben-David

Grading

The course grade points are distributed as follows for a total of 100%:

- **10-605:**
  - 5 Homework assignments (1x 5 + 4 x 10 = 45pts)
  - 1 Midterm Exam (15 pts)
  - 1 Final Exam (15pts)
  - 1 Scribing: (3pts)
  - Class Activities/Quizzes (5pts)

  Sum = 83pts

- **10-805:**
  - 5 Homework assignments (1x 5 + 4 x 10 = 45pts)
  - 1 Midterm Exam (15 pts)
  - 1 Final Exam (15pts)
  - 1 Scribing: (3pts)
  - Class Activities/Quizzes (5pts)
  - 1 Project (Midterm Report 4 + Final Report 8 + Class Presentation 10) =22pts

  Sum = 105pt

Scribing

- **Please make groups and sign up for scribing**
- **Use the provided template**
- **Finish the notes in a week**
- **Indicate your contributions**
- **Feel free to add more material (figures, texts, references) in addition to the class material if it helps understanding the lectures**

- **We need 8 people for scribing the Thursday lecture**

10-805 Project

- Form groups ASAP
- Optimal group size is 3. (Occasionally 4 can also be good)
- Proposal due: Oct 2nd, Tu, 1:30pm [You can submit sooner!]
- Deliveries: Proposal, Midterm report, Final report in NIPS style
- Group presentations on Nov 27, Tu, and Nov 29, Th, 1:30pm.
- Presentations will be peer graded
- Each team will have a mentor TA. Please meet you mentor TA biweekly.
- Topic: applying large scale machine learning approaches to your research area.
- Final report due: Dec 11, Tuesday. **NO LATE DAY!**

The ideal project is ambitious but reasonable. Plan such a way that if your project is successful it can be submitted to a top conference e.g. NIPS, ICML, CVPR, etc.

Homework Assignments

- Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early
- Homework questions are hard
- They have to be electronically submitted on Autolab!

Collaboration Policy

- You may **discuss** the questions
- **Each student writes their own answers**

  … copying from anywhere including whiteboard, emails, papers etc is not acceptable!

- **Each student must write their own code** for the programming part

  … simply renaming variables is not acceptable!

- **Please don't search for answers on the web, Google, previous years' homeworks, etc.**
  - please ask us if you are not sure if you can use a particular reference
- **Cheating:** Will be reported, will not be tolerated…

First Point of Contact for HWs

- To facilitate interaction, a TA will be assigned to each homework question.
- This will be your "first point of contact" for this question.
- If questions are not clear, you can discuss them on PIAZZA, but do not discuss the solutions there before the official solutions are released!…

Communication Channel

- Announcements will be posted on the class website and piazza.
- Please come to our office hours and ask questions.
- Come to recitations and ask questions.
- Ask questions at the beginning of the lectures.
- Use Piazza for discussions with other students.
- Email discussions with the Instructor might not work …

Feedback

Your feedback is highly appreciated!

If you have any concerns about the class (including homework, lectures, recitations, logistics, etc) please let us know using the google form on the class website. We appreciate any kind of feedback!

https://sites.google.com/site/scalableml2018f/feedback

Here you can leave anonymous messages for us. The Instructor and the TAs will discuss your comments during their weekly meetings.

Meetings with Barnabas

- Office hours

  Or

- Email Barnabas's assistant, Sharon Cavlovich: sharonw@andrew.cmu.edu in case of emergency.

Time Allocation

10605: 12 hours per week:

- 3h lectures
- 1h recitation
- 1.5h slides
- 1.5h reading material
- 5h homework

10805: 12 hours per week:

- 3h lectures
- 1h recitation
- 1.5h slides
- 1.5h reading material
- 5h projects + homework

The Team

Instructor:

Barnabas Poczos

- bapoczos@cs.cmu.edu
- office hours: Tu,Th after class, 2:50-3:30pm, PH100, then GHC 8231

Class Assistant:

Sharon Cavlowich

- sharonw@andrew.cmu.edu
- office: GHC 8221

The Team

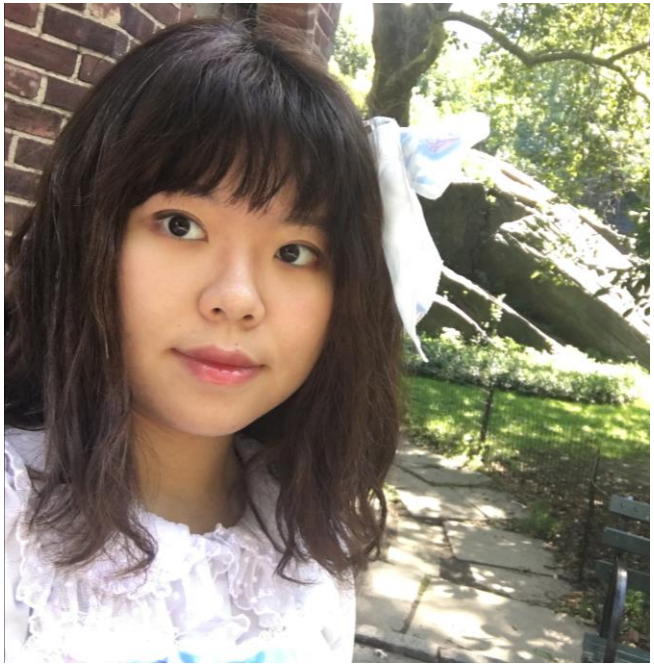
Teaching Assistants:

- Yiwen Yuan, yiweny@andrew.cmu.edu
- Tianxiong Wang, tianxiow@andrew.cmu.edu
- Yujie Ai, yujiea@andrew.cmu.edu
- Peixin Sun, peixins@andrew.cmu.edu
- Zheng Jiang, zjiang1@andrew.cmu.edu
- Ran Huan, rhuan@andrew.cmu.edu
- Yifan Wu, yw4@andrew.cmu.edu
- Harini Kesavamoorthy, hkesavam@andrew.cmu.edu

The Team

Yiwen Yuan

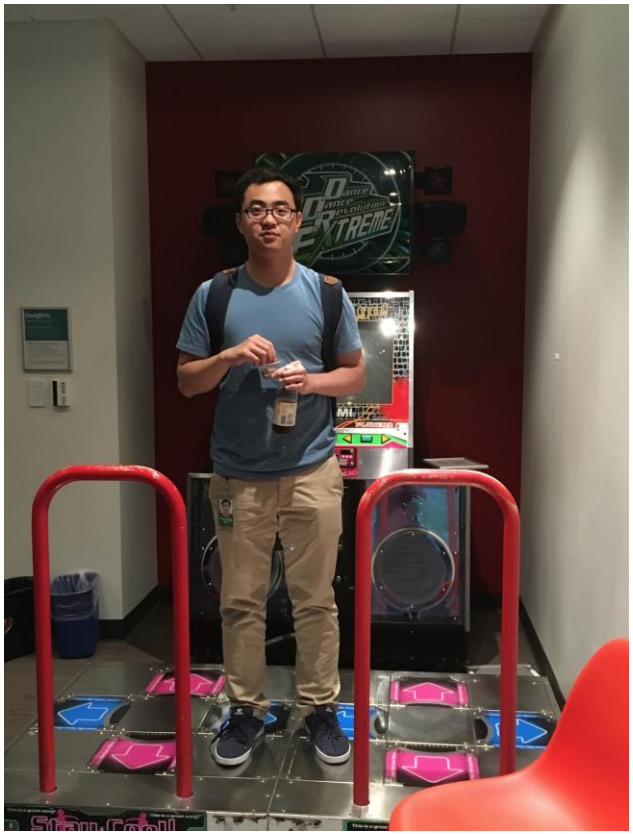
- Junior student studying Computer Science, planning to minor in Machine Learning
- Took this class last semester with Prof. William Cohen and loved it
- Taking 10-701 this semester
- Hobbies include drinking bubble tea with friends and traveling



The Team

Tianxiong Wang (MSCS)

- I am a second year master student under the computer science department
- I am very interested in large scale data processing and machine learning algorithms and learned a lot of useful topics from 10605 last fall.
- Some of the materials in the course were pretty useful during my internship at Google while working on recommendation system.



The Team

Yujie Ai (MSIN)

- Second year master student in Information Networking Institute
- Interested in Scalable Machine Learning and Computer Vision
- Took this course last semester and learnt a lot of interesting and creative ideas. Hope you enjoy it!

The Team

Peixin Sun

- I'm a second year master student from MCDS, SCS. During the summer, I interned at Uber.
- I took 605 last year and finished a course project about subgraph enumeration.
- Nice to meet you and hope you enjoy the course.

The Team

Zheng Jiang

- 2nd year of Master of Music & Technology
- Bachelor Degree of Software Engineer
- Interned in Cisco System and A9.com
- Nice to meet you!

The Team

Ran (Sharon) Huan

- Math undergrad. CS and stats ML additional majors.
- Took 10605 last year and enjoyed the course.
- Glad to work with you all!

The Team

Harini Kesavamoorthy

- I am a second year masters student in the Language Technologies Institute. I currently work on Code-Switching and Knowledge Acquisition for chatbots.
- I took 10-605 in the fall last year and had a great experience, learning interesting concepts.
- I hope you enjoy the course and learn a lot!

Any other questions about administration and logistics?

Todos

- Slides for TAs who haven’t sent it yet
- Set up AutoLab
- Set up Gradescope
- Create Piazza
- Create overleaf template for scribing for each lectures
- Google sheet for signing up for scribing
- Weekly meeting time
- HW1 test
- HW2 topics
- Edit rights for class website
- 1st recitation time and topic

The Team

- Yifan Wu:
- I am a PhD student from MLD.
- In general I am interested in fundamental challenges in pushing machine learning into practical use.
- My current research focuses on reinforcement learning.