10-718 Data analysis

Fall 2019, Syllabus

August 26, 2019

1 Basic Course Information

Instructor Leila Wehbe, lwehbe@cmu.edu [Office hours: on demand]
Assistant Instructor: Fabricio Flores, wflores@andrew.cmu.edu [Office hours: TBA]
Teaching Assistants:
Aria Y Wang, ariawang@cmu.edu [Office hours: TBA]
Jacob Tyo, jtyo@cs.cmu.edu [Office hours: TBA]

Time: Mon, Wed 4:30-5:50pm
Location: GHC 4307
Exceptions: Classes begin August 26th and end December 4th. No classes will be held on September 2nd (Labor day) and November 25th and 27th (November 27-29 is the Thanksgiving break). See the CMU academic calendar for the add-drop deadlines and other important dates.

Website See http://www.cs.cmu.edu/∼lwehbe/10718_F19 for course material.
Announcements All announcements will be made on Piazza, use the sign up link: piazza.com/cmu/fall2019/10718.

Participants This course is intended for MLD PhD and Masters students (including secondary Masters in ML).
Prerequisites There are no formal prerequisites for this course except a general knowledge of Machine Learning principles and an interest in how to apply them to real problems.

Textbook The purpose of the class is to generate a shared online textbook about case-studies.

As background material, some of the following textbooks might be useful:

- Exploratory Data Analysis by Tukey.
- How to lie with statistics by Huff.
- Advanced data analysis from an Elementary Point of View by Shalizi.
- The practice of reproducible research.
- Fairness and Machine Learning by Barocas, Hardt and Narayanan.
- Statistics gone wrong by Reinhart.

2 Course Description

In this course students will gain exposure to practical aspects of machine learning and statistical data analysis. Through a series of case studies of real problems, students will learn to appreciate the intricacies involved in the practical application of machine learning. The course will focus on formalizing research questions, data exploration,

identifying potential pitfalls, using machine learning for science and decision making, reproducibility and fairness. The outcome of the course will be a write up of the various case studies that will be shared between all students and possibly posted online (subject to agreement between students).

3 Graded components

There will be no exam in this class. Half of the grade will be based on group work. Each group will be assigned a week of the class and will have to:

- present material about that week's topic and lead a discussion between students (15%).
- produce a written document about that topic based on the readings and on the group discussion (35%) .

The second half of the grade will be based on individual work through participation in the class activities:

- Answering reading related quizzes weekly (20%), the grade for the 8 best quizzes (out of 11) will be counted.
- Filling out weekly suggestion sheets when other students present their blog post ideas (20%), the grade for the 8 best suggestion sheets (out of 10) will be counted.
- Providing helpful and respectful edits and review for one other blog post in that semester (10%).

3.1 Presentation, class discussion and written blog post

Each group of students will be responsible for one week's discussion and blog post (published internally to course participants and possibly externally). The group should present the relevant literature and lead the discussion (on the first class). On the second class, the group should present what they intend to write in the blog post, and solicit feedback from the class. The blog post should contain some combination of the following items (in any order), which can vary a bit depending on the topic:

- Diagram of the main points of the post (a graphical abstract).
- Explanation of the major concepts and goals.
- Detailed explanation of positive outcomes (with examples) and negative outcomes or potential pitfalls (with examples). Examples from student's own research are welcome.
- References to literature (class readings AND additional material). If readings are about a specific application, an explanation of what conclusions can be generalized to other applications is beneficial.
- Supporting plots with self-sufficient captions (and proper attribution).
- Discussion (including most interesting discussion points provided by the class, including disagreement, and suggested avenues for future work or improvement).
- Summarizing conclusion.

The first draft of these blog posts are due on the Tuesday following the specific week corresponding to the topic. The blog posts will have a final due date for grading, but they can be edited until the last day of the reading period before finals (December 10th). Each group should meet with the instructor at least 10 days before their first presentation.

3.2 Homework and readings

Each week, each student should read the material for the coming week (typically, two papers or a combination of a paper and newspaper articles).

3.3 Use of Mobile Devices and Laptops in Class

Using mobile devices or laptops in class is not allowed. Instead, you are encouraged to bring a printed-out version of the paper or your own paper notes to the class. Learning research shows that unexpected noises or movement automatically divert and capture people's attention, meaning that you are affecting everyone's learning experience. For this reason, I ask you to put away off your mobile devices and close your laptops during class. If you believe you have a valid reason for the use of a laptop, please reach out to one of the course staff.

This course is a discussion based course. A large portion of it will depend on in-class presence. If you feel that you need to use a screen during class because of lack of interest, this might mean you are not participating actively, and should focus instead on contributing to the discussion.

4 Learning Objectives

Upon successful completion of the class, students will be able to

- Formulate a research question and have a critical eye for projects or publications for which the goal is ill defined.
- List important reasons for visualizing and understanding data. Develop a set of possible steps and strategies for exploring a dataset.
- Explain the role of domain experts in defining the goal, understanding the data and interpreting the results.
- Define and develop a proper baseline.
- Evaluate the performance of a machine learning algorithm and be able to compare it with other models.
- Identify overfitting at various levels: not just in parameter fitting or hyperparameter tuning, but also in lack of generalization to new settings or in asking different questions based on the same data.
- Explain some of the intricacies involved in interpreting the results of machine learning algorithm to make conclusions.
- Explain the difference between prediction and causal claims.
- Explain several aspect of data science and statistical reproducibility.
- Explain multiple fairness issues inherent in today's machine learning approaches.
- Explain multiple negative societal impacts of today's machine learning approaches.

5 Approximate Schedule

This is a breakdown of the topics with a sample of the suggested readings. Every week, there will be two specific papers assigned to read (or one paper and one or a couple newspaper articles). The assigned papers will be communicated on Piazza. Students are encouraged to suggest other reading material or problems from their research.

- **Week -1 (Aug.26,28): Introduction to course.**
- **Week 0 (Sep.4): Introduction to course.**
- **Week 1 (Sep.9,11): Formulating a research question.**

  How to clearly define a data analysis goal? What is the notion of success? How is this different in a hypothesis testing setup or a prediction setup? What are possible fallacies to think of and what are strategies to prevent them? Proposed readings:

  - *Deconstructing Statistical Questions* by Hand.
  - *Intellectual Debt: With Great Power Comes Great Ignorance* by Zittrain.

- *Artificial Intelligence — The Revolution Hasn't Happened Yet* by Jordan.
- More examples of Simpson's Paradox on Wikipedia.
- *Sex Bias in Graduate Admissions: Data from Berkeley* by Bickel et al.

- **Week 2 (Sep.16,18): The importance of domain expertise and of understanding your data.**

  It is extremely important to understand where your data came from, how it was collected, what the different variables mean etc. Typically, many meetings with the domain experts including the people in charge of collecting the data will be necessary. This will vary by field, but typically, any data analysis result will not be impactful without understanding what it truly means. Any time spent understanding the data before diving in will likely avoid the need for lengthy future back and forth and model redesign. The readings in this week will be determined by the students in case they want to focus on their area of expertise. Alternatively, the instructor can suggest a specific dataset/associated papers.

- **Week 3 (Sep.23,25): Data exploration.**

  Before you use any kind of algorithm on your data, you need to look at it, understand it, plot summary statistics and distributions and identify important trends in it, missing data or major outliers. This step might allow you to identify features that are specific for that data that can help you achieve your goal easily. It can also help you determine what kind of algorithm to use and make decision such as whether a complex model (such as a deep model) is even necessary. The readings in this week will be determined by the students in case they want to focus on their area of expertise. Alternatively, the instructor can suggest a specific dataset/associated papers.

- **Week 4 (Sep.30,Oct.2): Evaluating classifiers, robustness**

  What does it mean for a model to predict well? How to evaluate a model performance, measure its reliability, and compare it to other models? We always want to build better models that what's available, but do we want methods that have drastically different performance? Proposed readings:

  - *Three principles of data science: predictability, computability, and stability (PCS)* by Yu and Kumbier.
  - *Statistical Comparison of Classifiers over Multiple Data Sets* by Demsar.
  - *Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations* by Stapor.
  - *The Secrets of Machine Learning: Ten Things You Wish You Had Known Earlier to be More Effective at Data Analysis* by Rudin.

- **Week 5 (Oct.7,9): Constructing a baseline**

  An inherent part of showing that a proposed machine learning model is good is to show that it performs better than what already exist, through the use of a baseline. How does one choose a good baseline? Further, is the comparison fair when one's model is tuned and optimized and the baseline model isn't? Proposed readings:

  - *Always start with a stupid model, no exceptions* by Ameisen.
  - *On the difficulty of evaluating baselines* by Rendle and Zhang.
  - *Beating State of the Art by Tuning Baselines* by Tatman.
  - *Stronger Baselines for Trustable Results in Neural Machine Translation* by Denkowski and Neubig.

- **Week 6 (Oct.14,16): Overfitting.**

  We learn in machine learning classes how to compute bounds on generalization error as well as general techniques to prevent models from overfitting. What are different ways that overfitting can happen, and how does one arrive at a machine learning result (even on test data that they didn't train on) that doesn't reflect true generalization? Do researchers truly stop after testing on the test set? What are the downsides from everyone using the same benchmark tests? The readings in this week will be determined by the students in case they want to focus on their area of expertise. Alternatively, the instructor can suggest a specific dataset/associated papers. One example could be an analysis of state of the art performance on NLP tasks: how well do these models generalize, do they actually solve the underlying task?

- **Week 7 (Oct.21,23): Methods and Results Reproducibility**

  Reproducibility is an overloaded words that people use to mean many things, such as the repeatability of a result when a scientific experiment or clinical trial is ran another time, or the ability of an experiment to be repeated in the first place through the sharing of data, code, correct description etc. These different notions of reproducibility are related to the correct use of statistical methods and experimental design, as well as other data science elements such as versioning code, creating data repositories and even clear writing. Proposed readings:

  - *What does research reproducibility mean?* by Goodman et al.
  - *Why Most Published Research Findings Are False* by Ioannidis.
  - Any chapter from *The practice of reproducible research.* edited by Kitzes et al.
  - *fMRI gets slap in the face with a dead fish* by NeuroSkeptic.

- **Week 8 (Oct.28,30): Model Interpretability**

  Whether we are using our model for scientific purposes or not, it is very interesting to know why it works the way it does. This is particularly interesting in the case of deep learning models which perform very well but that we don't understand theoretically. How can one correctly interpret how a model makes predictions, or what features of the input are used by different parts of the models? What are the fallacies to watch out for? Proposed readings:

  - *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* by Ribeiro et al.
  - *The mythos of model interpretability* by Lipton.
  - *Explanation in Artificial Intelligence: Insights from the Social Sciences* by Miller.
  - *The building blocks of interpretability* by Olah et al.
  - *Towards A Rigorous Science of Interpretable Machine Learning* by Doshi-Velez and Been Kim.
  - *This Looks Like That: Deep Learning for Interpretable Image Recognition* by Chen et al.
  - The Secrets of Machine Learning: Ten Things You Wish You Had Known Earlier to be More Effective at Data Analysis by Rudin.

- **Week 9 (Nov.4,6): Causality**

  Correlation doesn't imply causation, we all know that. However, what is the implication of this fact when building ML models that uniquely rely on finding patterns in offline data and without interventional experiments? How can we trust predictions from a model, and how do these predictions differ from a mechanistic understanding of the problem at hand? Proposed readings:

  - First chapter of *Elements of Causal Inference* by Peters et al.
  - *The Seven Tools of Causal Inference, with Reflections on Machine Learning* by Pearl.
  - *Explanation in Artificial Intelligence: Insights from the Social Sciences* by Miller.

- **Week 10 (Nov.11,13): Fairness and bias**

  ML models are trained on data that has inherent biases, and incorporate these biases. This has major implications when using ML models to make real world decisions. How does ML impact social fairness, and what research directions are being proposed as solutions? (This topic can be the focus of an entire course). Proposed readings:

  - *How Algorithms Can Bring Down Minorities' Credit Scores* by Waddell.
  - *Responses to Critiques on Machine Learning of Criminality Perceptions* by Wu and Zhang.
  - *Crime-prediction tool PredPol amplifies racially biased policing, study shows* by Smith.
  - *A tutorial on fairness in machine learning* by Zhong.

- *Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices* by Raghavan et al.

- **Week 11 (Nov.18,20): Accountability, transparency, Societal impacts**

  Many other aspect of ML are required in order to prevent it from causing other types of societal harm (additionally to bias). Again, entire courses can be devoted to these topics, but we will survey here some important aspects of accountability and transparency in the use of ML in real world application, as well as the role that ML has in impacting societal development. Proposed readings:

  - *Machine Learning for the developing world* by De-Arteaga et al.
  - *The case for technology in developing regions* by Brewer et al.
  - *The challenges of technology research for developing regions* by Brewer et al.
  - *Should I open source my model* by Munro.
  - *"Anonymized" data really isn't—and here's why not* by Anderson.
  - *Google's DeepMind made 'inexcusable' errors handling UK health data, says report* by Vincent.
  - *Here's What Happened When A Football Team Decided To Give Out Free DNA Tests* by Lee.

- **Week 12 (Dec.2,4):** Introduction / Conclusion

  In this week, we will revisit what we have built throughout the class, make joint decisions on how to format the blog posts and whether to release them, as well as draft together an introduction post and a conclusion post.

6 Course policies

6.1 Attendance

This course is a discussion based course and therefore attendance is essential for benefiting from it and contributing to it. Attendance will be taken in class and constitutes part of the grade. We understand that some events such as conferences will lead to absence. Please communicate your absence in advance to the course staff. There will be some allowance in the grade for a couple of absences.

6.2 Collaboration

Discussion of class material is heavily encouraged. Additionally,

- Direct plagiarism during the writing of blog posts will not be accepted. Material from sources can and should be used, but only when summarized and cited appropriately.
- Discussion is **not allowed** during quizzes.
- Discussion is **allowed** during the filling of suggestion forms.

6.3 Academic Integrity

We have a zero tolerance policy for violation of class policies. If you are in any doubt in regards to the policy, please clarify with the course staff before proceeding.

- Any deviation from the rules will be dealt with according to the severity of the case. For example: blindly copying one solution from someone else will result in the maximum points that can be earned for that quiz becoming zero (maximum eligible grade becomes B); repeat occurrences will result in a failing grade for the course.

- In line with university policy, all instances of cheating/plagiarism will be reported to your academic advisor and the dean of student affairs. See the university policy on academic integrity.

6.4 Late Assignments

The maximum earnable points for each assignment will drop by 20% per late day.

7 Additional information

7.1 Accommodations for Students with Disabilities

If you have a disability and are registered with the Office of Disability Resources, I encourage you to use their online system to notify me of your accommodations and discuss your needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

7.2 Statement of Support for Students' Health & Well-being

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night (CaPS: 412-268-2922, Resolve Crisis Network: 888-796-8226). If the situation is life threatening, call the police (On-campus CMU Police: 412-268-2323, Off-campus Police: 911).