Module 0: Introduction to Machine Learning and Data Mining (STA 325)

Professor Rebecca C. Steorts

Instructor

Prof. Rebecca Steorts (resteorts.github.io)

- Class: Wed/Fri, 11:45 - 1:00 PM EDT (Old Chemistry 116)
- OH: Wed/Fri, 1:00 - 2:00 PM EDT (Old Chemistry 216)
- Zoom ID: 953 6920 5290
- Course webpage: https://resteorts.github.io/teach/data-clean.html

Class will be in person unless you are notified! OH will be over zoom as students have tended to like this and I can answer questions more rapidly.

Teaching Assistants

- Jingcheng Meng, PhD Student
- Donald Cayton, MS Student
- Athena Ru, Undergraduate Student

Where to find information

- Course website (all major course information here): https://resteorts.github.io/teach/data-clean.html
- Course syllabus https://github.com/resteorts/data-clean/blob/main/syllabus/syllabus-sta325-fall24.pdf

Where to find information

- Course cheat sheet (a summary of course information): https://github.com/resteorts/modern-bayes/blob/master/syllabus/deadlines-cheatsheat.pdf
- Canvas: https://canvas.duke.edu/?login_success=1
- Duke Gradescope (upload homeworks): This should be directly accessible through Canvas.

Course resources

- Data Cleaning Pipeline: https://github.com/resteorts/data-clean/blob/main/articles/steorts-article-final-accepted.pdf

- (Almost) all of entity resolution: https://arxiv.org/pdf/2008.04443

- Four Generations of Entity Resolution https://link.springer.com/book/10.1007/978-3-031-01878-7

- An Introduction to Statistical Learning https://www.statlearning.com/

Please do read the first two articles above within a week and be prepared to discuss them.

Other resources

- Review of probability material: https://github.com/resteorts/modern-bayes/blob/master/reading/statistical-inference.pdf
- Simon Mak's Quick Guide to Prob. Distributions https://github.com/resteorts/modern-bayes/blob/master/reading/distribution-quick-reference.pdf
- A One Pager on Prob Distributions https://github.com/resteorts/modern-bayes/blob/master/reading/common-distributions-one-pager.pdf

Prior Knowledge

- STA 210 https://www2.stat.duke.edu/courses/Spring19/sta210.001/
- STA 230 https://www2.stat.duke.edu/courses/Fall18/sta230/
- Linear algebra http://www.stat.columbia.edu/~fwood/Teaching/w4315/Fall2009/lecture_12
- R programming (STA 199) https://www2.stat.duke.edu/courses/Spring18/Sta199/
- github (STA 199) https://www2.stat.duke.edu/courses/Spring18/Sta199/

Course Objectives

- Provide a foundation to statistical machine learning
- Understand the fundamentals of data cleaning and record linkage
- Understand data cleaning applications and their importance in machine learning
- Provide a basic understanding of data cleaning methods and algorithms
- Learn to effectively communicate methods/algorithms/results through written assignments and exams

Machine learning versus statistics

Statisticians often focus on:

- uncertainty quantification
- theoretical performance or proofs
- extending models that are well established
- applications to well known fields

Machine learners focus on:

- algorithms and computational improvements
- empirical performance on many data sets
- applications in tech and industry
- novelty methods or algorithms

What is the focus of this course?

- Modern data cleaning and machine learning methods.
- Such methods include data cleaning, data pipelines, entity resolution, record linkage, and thinking about problems in ways you may have not thought about before.
- We will not focus on classical ML methods, such as linear regression, logistic regression, bagging, boosting.
- Classical ML methods are taught in many courses at Duke, so take a different course if you want to learn these skills or read about them on your own.
- Again, do not expect to learn about classical ML models or you will be disappointed!

What is data cleaning?

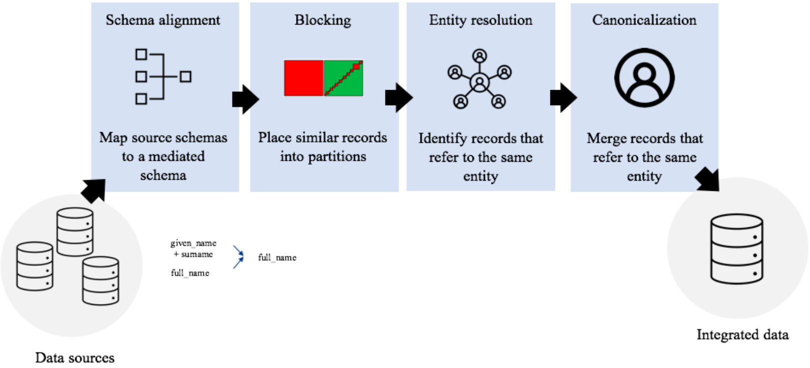

Schema alignment
Map source schemas to a mediated schema
Blocking
Place similar records into partitions
Entity resolution
Identify records that refer to the same entity
Canonicalization
Merge records that refer to the same entity
given_name + surname
full_name
full_name
Data sources
Integrated data

Figure 1: Data cleaning pipeline.

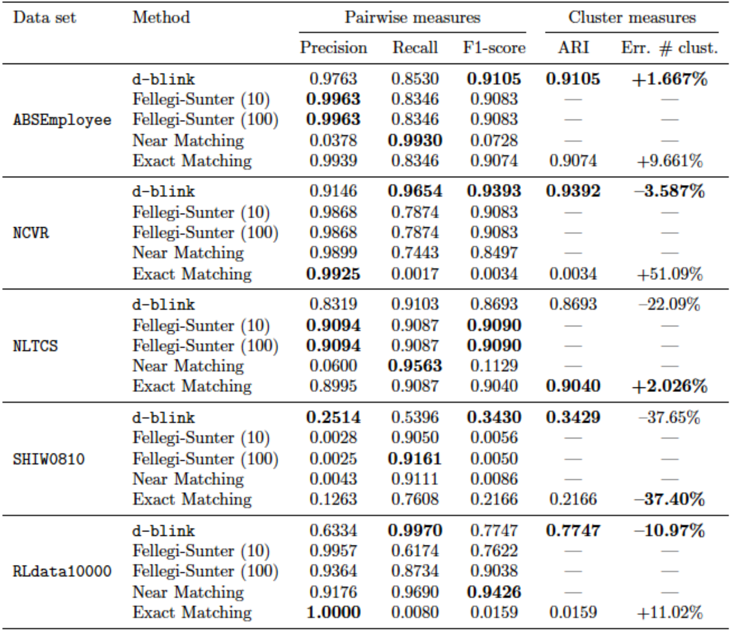
The scope of this course

- Cover important data cleaning methods
  - Examples include deterministic and probablistic methods. Others include supervised and unsupervised methods.
- Read core data cleaning papers and discuss them
- Be able to implement some data cleaning methods in R
- Be able to write about and discuss data cleaning methods
- Understand when certain methods are appropriate and when to utilize them over alternatives.

Is it possible for one method to always dominate?

- No method dominates all others, across all problems.
- For any two methods, each one will perform better on some problems, comparatively.
- For evidence, refer to Wolpert (1996) as the “no free lunch problem.”
- However, some methods to often perform better than others on certain types of application areas or data sets.
- What do students think of this idea? It is intuitive or confusing?

What do you observe?



Data set	Method	Pairwise measures			Cluster measures	
		Precision	Recall	F1-score	ARI	Err. # clust.
ABSEmployee	d-blink	0.9763	0.8530	**0.9105**	**0.9105**	**+1.667%**
	Fellegi-Sunter (10)	**0.9963**	0.8346	0.9083	—	—
	Fellegi-Sunter (100)	**0.9963**	0.8346	0.9083	—	—
	Near Matching	0.0378	**0.9930**	0.0728	—	—
	Exact Matching	0.9939	0.8346	0.9074	0.9074	+9.661%
NCVR	d-blink	0.9146	**0.9654**	**0.9393**	**0.9392**	**–3.587%**
	Fellegi-Sunter (10)	0.9868	0.7874	0.9083	—	—
	Fellegi-Sunter (100)	0.9868	0.7874	0.9083	—	—
	Near Matching	0.9899	0.7443	0.8497	—	—
	Exact Matching	**0.9925**	0.0017	0.0034	0.0034	+51.09%
NLTCS	d-blink	0.8319	0.9103	0.8693	0.8693	–22.09%
	Fellegi-Sunter (10)	**0.9094**	0.9087	**0.9090**	—	—
	Fellegi-Sunter (100)	**0.9094**	0.9087	**0.9090**	—	—
	Near Matching	0.0600	**0.9563**	0.1129	—	—
	Exact Matching	0.8995	0.9087	0.9040	**0.9040**	**+2.026%**
SHIW0810	d-blink	**0.2514**	0.5396	**0.3430**	**0.3429**	–37.65%
	Fellegi-Sunter (10)	0.0028	0.9050	0.0056	—	—
	Fellegi-Sunter (100)	0.0025	**0.9161**	0.0050	—	—
	Near Matching	0.0043	0.9111	0.0086	—	—
	Exact Matching	0.1263	0.7608	0.2166	0.2166	**–37.40%**
RLdata10000	d-blink	0.6334	**0.9970**	0.7747	**0.7747**	**–10.97%**
	Fellegi-Sunter (10)	0.9957	0.6174	0.7622	—	—
	Fellegi-Sunter (100)	0.9364	0.8734	0.9038	—	—
	Near Matching	0.9176	0.9690	**0.9426**	—	—
	Exact Matching	**1.0000**	0.0080	0.0159	0.0159	+11.02%

Missed class/review material?

Go to https://github.com/resteorts/modern-bayes/ and fork the repository.

- Make sure to pull the repository each day. I update the repository very often as all the course resources are here (homeworks, lectures, data). If you don't pull often, you might run into some issues!

Class Meetings

- Learn about statistical machine learning
- Lectures will consist of learning methodology and applied coding techniques
- You might find it useful to have a tablet/document camera for homeworks/OH/exams
- An alternative to this is using your phone to take pictures using Evernote (students have said this worked well in the spring). Please test things out in advance and make sure things are legible!

Labs

- Lab will be used as an opportunity to work with teaching assistants on course content and ask questions

Activities & Assessments

- Homework: Individual assignments combining conceptual and computational skills along with lab exercises. *Lowest score dropped.*
- Labs: **Attending lab is highly encouraged** but not required. The goal is to learn about modern statistical machine learning methods.
- Class/Lab: **You will be responsible for keeping up with all class, reading, and homework material on a weekly basis.

Class Engagement

0. Students are expected to have read any assigned reading before coming to class.
1. Professor Steorts will prepare slides regarding material and go through these during class.
2. Students are encouraged to ask questions and engage in class.
3. The purpose of homeworks and exams will be to test the understanding of the course knowledge from the course material.

Homeworks

1. All code must be written to be reproducible in Markdown.
2. All derivations can be done in any format of your choosing (word, latex, markdown, written by hand) but must be converted to a pdf document. It must be legible.
3. You reproducible Rmd file is uploaded to Canvas. Your PDF file is uploaded to Gradescope.
4. Ask questions early if you have a problem to a TA regarding submission issues.
5. Your lowest homework will be dropped.

Remark: Canvas is for reproducible code. Gradescope is to make grading easier for everyone. Unfortunately, there is not a platform that handles both.

Please see the syllabus for all homework guidelines.

Why Canvas + Gradescope?

1. Canvas will be used for reproducibility so this can be checked. This will account for part of your grade each assignment.
2. Gradescope will be used to:

- return homework/exams more promptly
- ensure complete fairness
- allow students to continue to learn from homeworks by reworking problems without the right answer written out to the side

Exams

The format of the exams will be discussed closer to the time of the exam. At this time, plan to be for the exams in person unless there is an announcement about this.

Grade Calculation

Component	Weight
Homework	30%
Exam 1	20%
Exam 2	20%
Exam 3	30%

- See the syllabus for grade breakdowns.
- Grades will **never** be curved down.
- You are expected to attend lectures and labs in order to keep up with the course material.
- There will be no attendance grade or participation grade.

Absences

- Students who miss a class due to a scheduled varsity trip, religious holiday, or short-term illness should fill out the respective form.
  - These excused absences do not excuse you from assigned work.
- If you have a personal or family emergency or chronic health condition that affects your ability to participate in class, please contact myself and or your academic dean's office.
- Exam dates cannot be changed and no make-up exams will be given.
- What if I have COVID symptoms?
  - Please do not come to class or return to class until you are cleared to come back. Forms can be found here: https://trinity.duke.edu/undergraduate/academic-policies/illness

Late Work and Regrade Requests

- No late homeworks will be accepted, so please do not ask.
- No make up exams will be given.
- Regrade requests must be submitted within one week of when the assignment was returned.

Academic Honesty

All work for this class should be done in accordance with the Duke Community Standard.

To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised. Any violations will automatically result in a grade of 0 on the assignment and will be reported to Office of Student Conduct for further action.

Reusing Code

- Unless explicitly stated otherwise, you may make use of online resources (e.g. StackOverflow) for coding examples on assignments. If you directly use code from an outside source (or use it as inspiration), you must or explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.
- On individual assignments, you may discuss the assignment with one another; however, you may not directly share code or write up with other students.
- On team assignments, you may not directly share code or write up with another team. Unauthorized sharing of the code or write up will be considered a violation for all students involved.

Where to find help

- **If you have a question during lecture or lab, feel free to ask it!** There are likely other students with the same question, so by asking you will create a learning opportunity for everyone.
- **Office Hours**: A lot of questions are most effectively answered in office hours, so please take advantage of these.
- Canvas: Outside of class and office hours, any general questions about course content or assignments should be posted on the forum since there are likely other students with the same questions. **Please be careful not to give away the answers to homework questions and please keep it polite/friendly on the forum.**

Academic Resource Center

Sometimes you may need help with the class that is beyond what can be provided by the teaching team. In that instance, I encourage you to visit the Academic Resource Center.

The Academic Resource Center (ARC) offers free services to all students during their undergraduate careers at Duke. Services include Learning Consultations, Peer Tutoring and Study Groups, ADHD/LD Coaching, Outreach Workshops, and more. Because learning is a process unique to every individual, they work with each student to discover and develop their own academic strategy for success at Duke. Contact the ARC to schedule an appointment. Undergraduates in any year, studying any discipline can benefit! Contact arc@duke.edu 919-684-5917, 211 Academic Advising Center Building, East Campus – behind Marketplace.

Technology/Other

- Make sure that you have your zoom ids organized so that you're not late for class.
- Ensure the volume on all devices is set to mute.
- Refrain from engaging in activities not related to the class discussion. Browsing the web and social media, excessive messaging, playing games, etc. is not only a distraction for you but is also a distraction for everyone around you.
- If you have a question, I don't mind if you interrupt me during class.
- If you find a **typo in the slides**, please write these down and post these privately to the instructors on Canvas so they can be fixed.

Accessibility

Please contact the Student Disability Access Office (SDAO) if there is an element of the course that is not accessible to you. There you can engage in a confidential conversation about the process for requesting reasonable accommodations.

Please note that accommodations are not provided retroactively, so please contact them as soon as possible. More information can be found online at access.duke.edu.

Inclusion

In this course, we will strive to create a learning environment that is welcoming to all students and that is in alignment with Duke's Commitment to Diversity and Inclusion. If there is any aspect of the class that is not welcoming or accessible to you, please let me know immediately.

In addition, if you are experiencing something outside of class that is affecting your performance in the course, please feel free to talk with me and/or your academic dean.

Questions

Any questions regarding the format of the semester or any concerns?

Announcements

- **Please see me if you are on the waiting list**
- **Please do not come to class if you believe that you may be sick. Please reach out an let the TA team know what is going on so that we can know what is going on and can be helpful.**
- **If your situation changes during the semester for any reason, please email myself and the TAs so that we can help you.**
- **In return, I would ask that everyone be flexible and understanding of everyone else (including instructors and TAs) as this is still a very difficult and trying time for everyone.**