Welcome to SML 310

Wassily Kandinsky, *Bustling Aquarelle* (1923)

SML 310: Research Projects in Data Science, Fall 2019

Michael Guerzhoy
About me

• Michael Guerzhoy (pronounced “ger-JOY”)
• Started as a lecturer at CSML last year
• Working on data science for healthcare and on data science education
• Before that, some machine learning, some computer vision, some teaching, and some data science consulting
About you
About the class

• The goal is to support you in working on an interesting project in data science

• Lectures and mini-projects are meant to provide the knowledge and practical skills needed to get started with modern data science techniques
  • Python
  • Hierarchical models; Stan (a framework for fitting hierarchical models)
  • The basics of machine learning and neural networks; PyTorch (a framework for training neural networks)
About the class

• Initial project proposal
  • Requirements are on the website
  • Basically
    • What is the problem you are trying to address?
    • What kind of results might you expect and why? (Reference existing work)
    • Summarize the results and methods of at least two papers that addressed similar problems
  • It is understood that you may change your mind about what to work on (perhaps based on our feedback)
Course project

• Solve (or make progress toward solving, or produce a substantive negative result for) a problem using data science

• Many possibilities:
  • Collect or find an interesting new dataset that hasn’t been used before, and apply interesting data science techniques to it
  • Apply a method to your dataset that hasn’t been applied to that kind of dataset in exactly the same way
  • Obtain new insights about a dataset
  • Devise and use a new method, test it out on your dataset

• You don’t *have* to apply machine learning to a large-scale dataset
  • But running standard linear regression on 20 datapoints is unlikely to be approved as a project plan, even if the datapoints are really interesting
Course project: priorities

• Use deep (social-)scientific insight and data science to produce new, important, and interesting knowledge about the world
  • Publish, become famous, donate Nobel prize money to your alma mater
• Work on something interesting that you are excited about
• Produce something publishable
• Produce something useful (but complementary) to your thesis
• Do a nice project and get an A
Course project: examples

- Papers posted on the course website
- Some project ideas posted on Piazza
Course project: requirements

• Exploratory data analysis
• Overview of prior work
• Technical description of the data science method
• A description of how what you are doing relates to prior work
• Results
• Conclusions
Course project: grading

• This is a small class
  • I am not grading you on a curve (i.e., there is no quota for the number of A’s)
• You will be graded on the quality of the write-up, the quality of your ideas for the project, and on the work you will have done running experiments and/or collecting data
• The usual Princeton policy for A+’s: possible for extraordinary work but unusual
Course project: timeline

• Project proposal due Sept. 30
  • Read and summarize a few related papers, make sure you can get the appropriate dataset, make a plan
• Revised proposal due Nov. 12
  • Read and summarize related papers, do exploratory data analysis, write up a plan for the main analysis
• Presentation soon after the revised proposal
  • Present and discuss the problem you’re addressing
• Project write-up due on Dean’s date
  • Build on the revised proposal to complete the project
• Multiple deadlines, but really one piece of work – we’re just trying to keep you on track
Course project: research

• You are trying something new – can’t predict what you will discover or which direction you will take ahead of time

• It’s fine to change up your plans
  • But we really want to have something by Dean’s date (and by the other deadlines)
  • Talk to me if you need to change your plans
  • Sometimes it makes sense to just finish up what you started even if there’s a better idea out there
Mini-Projects

• Four mini-projects
  • MP1: statistical inference and hierarchical models
  • MP2: Python warm-up and building classifiers (+ data representation)
  • MP3: natural language processing
  • MP4: PyTorch and image data
Structure of the class

- First several weeks: Python workshops in precept
  - Python work sessions outside of class time
  - Attendance for Python precepts is not mandatory
- Rest of the semester: lecture + precept (+ presentations week)
  - Course grade includes participation
  - Don’t skip more than 15% of classes, do work during precept
Just the second offering of SML310

• Want to accommodate students with a variety of backgrounds, and varied amounts of experience in data science
• Your feedback is important!