Welcome to SML 310

Wassily Kandinsky, *Bustling Aquarelle* (1923)

SML 310: Research Projects in Data Science, Fall 2018
Michael Guerzhoy
About me

• Michael Guerzhoy (pronounced “ger-JOY”)
• Started as a new lecturer at CSML at Princeton eight days ago
  • After driving down from Toronto
• Since last year, working in machine learning for healthcare
  • Mostly working on building an early warning system to predict adverse outcomes for General Internal Medicine (GIM) patients
  • Still working with St. Michael’s Hospital in Toronto
• Before that, some machine learning, some computer vision, a fair bit of teaching, and some statistical consulting
About the class

• Project-based
• 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
  • The basics of machine learning and neural networks; PyTorch (a framework for defining and training neural networks)
  • Hierarchical models; Stan (a framework for fitting hierarchical models)
About the class

Mini-Projects

- **Mini-Project 0: Python warm-up (5%)** Due Oct 1 at 11PM
- **Mini-Project 1: NLP (8%)** Due Oct. 15 at 11PM
- **Mini-Project 2: Image data and PyTorch (8%)** Nov. 5 at 11PM
- **Mini-Project 3: Hierarchical Models (8%)** Due Nov. 19 at 11PM

Course project

- **Initial project proposal (1%)** Due Sept. 24 at 11PM
- **Revised project proposal (15%)** Due Nov. 12 at 11PM
- **Project presentation (10%)** to be scheduled during November and December
- **Course project (40%)** Due on the Dean’s Date at 11PM

• The Python assignment
  • Tentatively scheduled due Oct. 1
  • To be posted by Monday
  • We will discuss catching up on Python in the last 35 minutes of the class
About the class

• Initial project proposal
  • Will post requirements by the end of Friday
  • Basically
    • What is the problem you are trying to address?
    • What kind of results might you expect and why? (Reference existing work)
    • Briefly 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)
The course project

• Solve (or make progress toward solving, or produce a substantive negative result) 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
The course project

- Some mandatory aspects
  - 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
The course project: grading

• This is a small class
  • I am not grading you on a curve
• 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 course project

• The revised proposal (due Nov. 12) will contain a review of prior work, and a concrete plan for the course project
Structure of the class

• First two-thirds of the semester: lectures + occasional workshops (during precept time, but also outside of precept time)

• Last one-third of the semester: lectures on topics of interest to the class + project presentations
This is the first offering of SML310

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