ITAO40570 Urban Analytics

Machine Learing in Urban Analysis

Jeff Cai
Spring 2024

E-mail: jcai2@nd.edu
Office: MCOB 356
Class Room: MCOB L003

Web: piazza.com/nd/spring2024/itao40570
Office Hours: TR 3:30-4:30pm @ MCOB 356 or by appointment
Class Hours: TR: 12:30 - 1:45pm; 2:00 - 3:15pm

Urban analytics

Course Description

Urban regions will experience most future population growth, bringing opportunities and challenges. At the same time, statistical/machine learning has been evolving rapidly in the era of big data and provides tools to inform both data-driven decision-making and long-term planning in complex systems such as cities. Focusing on methodologies with statistical reasoning, the course brings in a large set of cutting-edge machine learning techniques combined with up-to-date urban case studies. We will cover data science essentials starting from data acquisition, exploratory data analysis (EDA), and visualization along with tools for reproducible reports. We next show how to build and interpret basic models; then we go beyond and focus on contemporary methods and techniques for handling large and complex urban data. While this course extensively uses the statistical programming language R, no programming experience is required. By the end of the semester, students will master popular modern statistical methods, but also get equipped with hands-on skills in urban data analytics.

Lecture notes will be provided by the instructor. They are organized by topics (modules) and written in the reproducible RMarkdown format which combines description, visualizations, explanation, and R codes.

Methods covered (mostly)

Part I: Acquiring, preparing, exploring and visualizing data

- R/Rstudio/Knitr
- Study design and data acquisition/preparation
- Exploratory Data Analysis (EDA)
- Spatial data: maps and flows
- Principal Components Analysis (PCA)
- Robust PCA
- Clustering
- Missing data

Part II: Model-based supervised learning

- Multiple regression
- Spatial regression
- Robust standard error estimation
- Training and testing errors
- k-fold cross validation
- Bootstrap
- Penalized regression: LASSO, Ridge Regression, Elastic Net
- Maximum likelihood estimator (MLE)
- Logistic Regression/Multinomial regression
- Classification/ROC/AUC and FDR

Part III: Machine learning

- K-nearest neighbors (KNN)
- Tree based methods (Bagging, Random Forest and Boosting)
- Neural network/Deep learning
- Text mining/Nature Language Processing (NLP)
- Network model
- Matrix completion (Recommendation system)

Case study/Datasets

- House price prediction
- Impact of AirBnB on rental market
- Real estate crowd-sourcing
- Chicago housing market
- Bike share system
- COVID-19: lock-down and compliance
- Can we do something to reduce crime rates?
- Using random forest to predict random forest fire
- Using Yelp reviews to predict the rating (text mining)
- Seeking alpha in REITs
- Fannie Mae loan performance

And more!

Course Materials

Software

The free and open source statistical computing language R is used through RStudio. There are infinitely many new packages available for us to use; an interface to explore the publicly available R packages is available via Microsoft. We will use RMarkdown for all materials to ensue reproducibility. We will also use Git and GitHub for version control and collaboration.

Throughout of the semester, we use the free RStudio, an interface for writing R documents and working with data.

Install the following software: R, RStudio, RMarkdown and git. Detailed instructions are available on Canvas.

R tutorial

- Basic R tutorial
  - Canvas: Get_staRted.Rmd/Get_staRted.html
- Advanced R tutorial
  - Canvas: advanced_R_tutorial.Rmd
  - covering `dplyr`, `data.table` and `ggplot`

Lecture notes

Over the years we have been developing our own lecture notes. They are organized by topic and written in reproducible RMarkdown format which combines R codes, visualizations, and narrative text. Real case studies are deployed throughout. The methods are explained with insightful ideas with minimum mathematics. Some deeper explanations and useful materials are postponed in Appendices as references. Students are urged to read through before classes and put hands on line by line at some point.

Textbooks

While we suggest you to read through thoroughly our own lecture notes which often cover more materials, we require you to have the following two books:

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning with Application in R (ISLR)*, Available freely online, Second Edition, 2021, Springer New York.
- Garrett Grolemund & Hadley Wickham, *R for Data Science*, 2016, O'Reilly. Available freely online.

An advanced text book as a reference:

- Trever Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, 2008, Springer. A PDF version of ESLR is available from the authors.

A reference for general statistics method you may read:

- Ramsey and Schafer, *The Statistical Sleuth*, Third Edition, 2013, Brooks/Cole (an e-version is available in the canvas site)

Course Policies

Communication

Communication will be through Canvas and Piazza. Course material will be uploaded to Canvas, including dataset, homework, and lecture notes. We will use Piazza as the Q&A platform.

I have been really enjoying teaching this class and care about your progress. Please do not hesitate to ask any questions or raise any concerns. I can definitely meet outside of office hours by appointment. I would be also happy to chat about the course or to help with career advice and grad school questions.

Laptop Policy

A **laptop** is a must for the course. You are encouraged to bring the laptop to classes so that you may run the lecture code simultaneously with the professor.

Assignments and Exams

1. **Homework (30%)**: We will give 4 or 5 homework assignments. These can be done in groups of up to 3 people; see the Group Policy for more details.
2. **Quizzes (15%)**: We will give two 10-minute in-class individual quizzes. The final quiz will be 40-minute as a final exam. *No makeup quizzes.* Please let me know in advance for special circumstances.

   ```
   * Quiz 1: 2/13 (Tue)
   * Quiz 2: 4/4 (Tue)
   * Quiz 3: 4/27 (Thu)
   ```

3. **Midterm (30%): 3/7 (Thu)** This exam will be an in class, *individual*, open-book and done on the computer. You will be given an exam in RMarkdown format to work through. Previous exams will be available on Canvas.
4. **Final Project (25%): 4/30 (Tue) presentation & 5/5 (Sun) report:** The ultimate goal of the class is to prepare/expose students to techniques that are suitable for modern data. The final project is designed so that each of you will bring a problem of personal interest to the class. You will need to identify a problem, collect/extract or find an appropriate data set, run a complete data driven study and make a final conclusion from your study.

This project is done with a group of up to 3 members. A complete write up is required. This would be a good project to put in your CV.

- A well-motivated, relevant topic is most desirable.
- Originality, complexity, and challenge will be another plus.
- A complete write up is a must.
- Setup a GitHub Page for the project (optional).
- **Maximum of 15 pages**.
- Some data sources:
  - US government's open data
  - US Census Bureau
  - Google provides public dataset through BigQuery on Google Cloud Platform.
  - NYC OpenData
  - gapminder
  - UCI Machine Learning Repo
  - Kaggle
  - Awesome Public Datasets

Late Work Policy

It is imperative that you manage your workload properly for this course. We will allow late assignments up to 3 days, with a 15% penalty per day. Note that lateness will be determined by the timestamp on Canvas submissions, i.e. 12:01 AM is considered late.

Group Policy

The homework and the final project can be done by groups of up to three people (can be from either sections). Sign up for groups on Canvas as soon as possible but no later than **Thursday, 1/25**. We will help out for those who need to find a group, with searches on Piazza.

Please note that at no time may a group have more than 3 members. In addition, while those within a group will submit a single homework file for the group, students must follow the code of academic integrity in regards to classmates outside their group. Finally, students do not have to complete the final project in the same group as for homework. They may form a new group though again no more than 3 people may be in a group. We prefer you keep the same groups through the semester but it is now required.

Grading Policy

- Homework: 30%
- Quizzes: 15% (6% for quiz 1 & 2; 9% for quiz 3)
- Midterm Exam: 30%
- Final Project: 25%

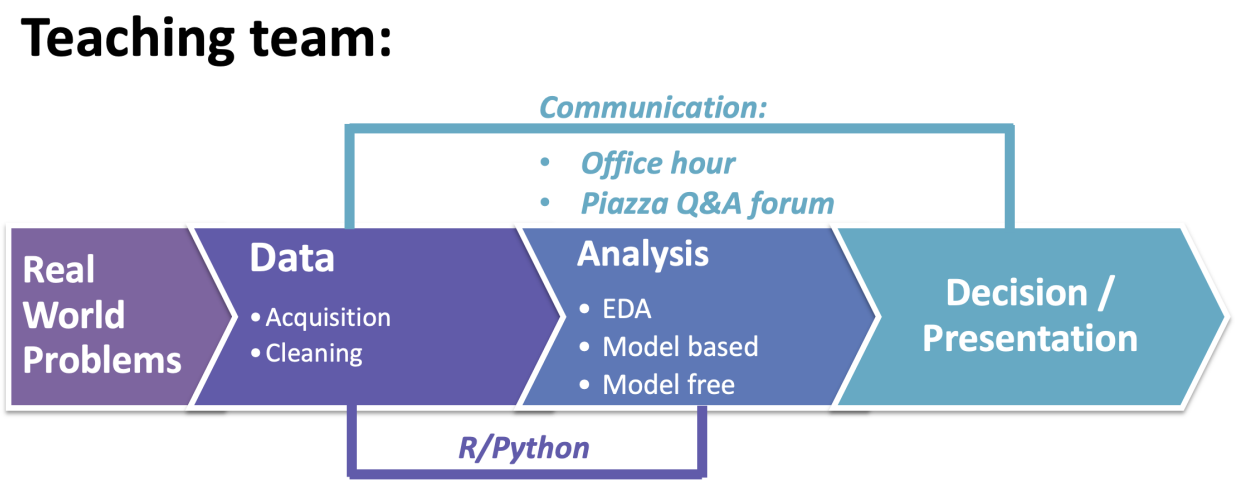
Academic Code of Honor

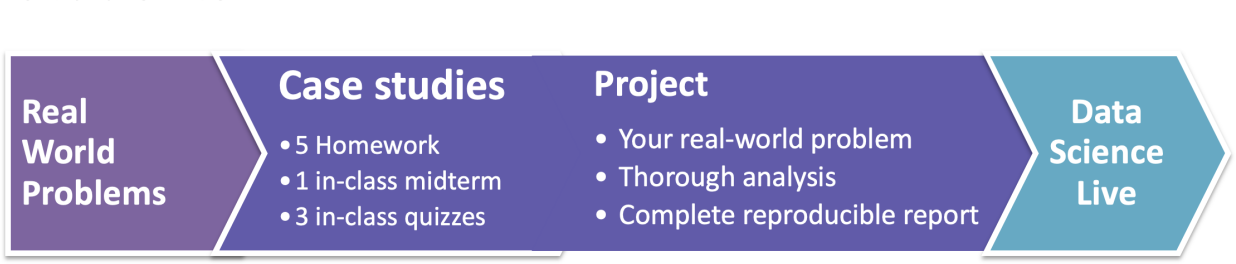
Notre Dame students are expected to abide by the Academic Code of Honor. "As a member of the Notre Dame community, I will not participate in or tolerate academic dishonesty." The graded work you do in this class must be your own, with the exception of assignments specifically designed and described as group work. In the cases when you collaborate with other students, make sure to fairly attribute their contribution to your project.

Sara Bea Accessibility Services

Any student who has physical or learning disabilities should speak with me as soon as possible and I will strive to make appropriate commendations. Also, you are encouraged to register with the Office of Disability Services.

Course Summary





Class Schedule

Tentative and subject to change. Unless otherwise noted, readings refer to *Introduction to Statistical Learning*.

Week 01, 01/15 - 01/21:

- **Tue 1/16**: Syllabus/Data acquisition and preparation
- **Tue 1/16**: Basic R Tutorial, 3:30 - 4:30 @ MCOB L003 (Optional)
- **Thu 1/18**: Exploratory data analysis (EDA)
- **Thu 1/18**: Advanced R Tutorial (dplyr/data.table/ggplot), 3:30 - 5:00 @ MCOB L003 (Optional)

Week 02, 01/22 - 01/28:

- **Tue 1/23**: Spatial data
- **Tue 1/23**: **Last date for class changes**
- **Thu 1/25**: Spatial networks
- **Thu 1/25**: **Grouping due on Canvas**

Week 03, 01/29 - 02/04:

- **Tue 1/30**: Dimension Reduction/Principal Component Analysis (PCA), (Ch 6.3.1, 12.1, 12.2)
- **Thu 2/1**: Continued topics

Week 04, 02/05 - 02/11:

- **Mon 2/5**: **Homework 1 due**, before 11:59 PM to Canvas
- **Tue 2/6**: Leeway
- **Thu 2/8**: Simple linear regression (Ch 3.1 - 3.6)

Week 05, 02/12 - 02/18:

- **Tue 2/13**: **Quiz 1**. Continued topics
- **Thu 2/15**: Leeway

Week 06, 02/19 - 02/25:

- **Mon 2/19**: **Homework 2 due**, before 11:59 PM to Canvas.
- **Tue 2/20**: Multiple linear regression
- **Thu 2/22**: Continued topics

Week 07, 02/26 - 03/03:

- **Tue 2/27**: K-fold Cross Validation (Ch 5.1.3) / LASSO (Ch 6.2)
- **Thu 2/29**: Continued topics

Week 08, 03/04 - 03/10:

- **Mon 3/4**: **Homework 3 due**, before 11:59 PM to Canvas.
- **Tue 3/5**: Review
- **Thu 3/7**: **Midterm**

Week 09, 03/11 - 03/17:

- **Tue 3/12**: Mid-Term break
- **Thu 3/14**: Mid-Term break

Week 10, 03/18 - 03/24:

- **Tue 3/19**: Spatial regression
- **Thu 3/21**: Continued topics
- **Fri 3/22**: **Last day for course discontinuance**

Week 11, 03/25 - 03/31:

- **Tue 3/26**: Spatial regression / Maximum likelihood estimator (MLE) (Ch 4.1-4.3.4)
- **Thu 3/28**: Logistic regression, Classification (ROC, AUC, FDR). Bayes rule

Week 12, 04/01 - 04/07:

- **Tue 4/2**: Continued topics
- **Thu 4/4**: Text mining

Week 13, 04/08 - 04/14:

- **Mon 4/8**: **Homework 4 due**, before 11:59 PM to Canvas.
- **Tue 4/9**: **Quiz 2**. Neural Network/Deep Learning
- **Thu 4/11**: Continued topics

Week 14, 04/15 - 04/21:

- **Tue 4/16**: Decision trees (Ch 8.1)
- **Thu 4/18**: Bagging/Random Forest (Ch 8.2)

Week 15, 04/22 - 04/28:

- **Mon 4/22**: **Homework 5 due**, before 11:59 PM to Canvas.
- **Tue 4/23**: Continued topics
- **Thu 4/25**: **Quiz 3**. Boosting (Ch 8.2)

Week 16, 04/29 - 05/05:

- **Tue 4/30**: **Final presentation**
- **Sun 5/5**: **Final project due before 11:59 PM to Canvas**