URPS 22: Undergraduate Research Program in Statistics for Winter 2022

Director of Undergraduate Programs: Prof. Edward Ionides

Undergraduate Program Coordinator: Gina Cornacchia

- Is this program for me?
- The application process.
- The projects.
- What to expect if I join a project.
- Do I get course credit? Can I use the project for an honors thesis? For the Data Science major capstone requirement?
- Other opportunities for undergraduate research in statistics.

1. Curriculum learning for reinforcement learning

Curriculum Learning (CL) is a strategy that has been widely applied in modern machine learning. It is built on the intuition that machines can learn faster when training samples are presented in a certain order, usually from simpler to harder. This fascinating idea is rooted deeply in the way humans learn, e.g., we learn from simpler skills that can be generalized to more difficult tasks. CL has also been shown to improve the performance of Reinforcement Learning (RL) in various applications including navigation, robotic control, and video games. However, we currently lack a proper understanding of why CL helps the training of RL algorithms. The goal of this project is to develop such an understanding. We are especially interested in how CL eases the difficulty of exploration, a main challenge of modern RL. The project involves (a) proposing meaningful simulation environments that can reveal the underlying mechanism whereby CL helps RL; (b) implementing deep RL algorithms in Python and running them on UM's Great Lakes cluster. Knowledge of Python and prior/concurrent enrollment in a machine learning (of any sort, deep or otherwise) course is a must. Familiarity with a deep learning library such as Pytorch or Tensorflow will be a plus.

Supervisors: Prof. Ambuj Tewari and Ziping Xu

2. Testing fairness deviations from fair aware datasets

Curating fair datasets is often the first step in the training pipeline when fairness among groups is of concern. Several papers present varying methods for optimizing these fair datasets, typically focusing on two methods. 1) Adaptive sampling to optimize for predefined fairness, classifier, and loss function. 2) Sequential sampling for minimax fair classification of a predefined loss function over a family of classifiers. It is known that achieving one type of fairness does not give guarantees for other types of fairness. Because of this, overspecification of fairness when curating datasets could become an issue when classifiers or loss functions are changed as is often the case in practice. This project will take datasets common in the fairness literature and curate fair datasets and train classifiers using methods from recent papers. We will then compare their results on different fairness objectives and classification methods to those using randomly sampled and balanced datasets. The goal is to look at whether the adaptive sampling methods perform significantly worse when used outside their original goal compared to these more naive datasets. Proficiency in basic Python is required, and some experience with pandas and sklearn.

Supervisors: Prof. Ambuj Tewari and Laura Niss

3. Mining genomic data

Many scientists and clinicians are interested in mining large-scale genomic datasets in order to better understand the relationships between gene expression patterns (more generally, human genomic markers) and disease outcomes (e.g. cancer/no cancer). One major statistical challenge presented by these data is that they often contain statistical patterns due to variables other than the disease outcome (e.g., age, sex). Failing to account for these other variables can lead to problems when we try to analyze the data. For instance, we might partially mistake the signal due to age for the signal due to the disease outcome, which can lead to incorrect inference and poor predictions. In this project, we focus specifically on classification analysis (e.g., using gene expression patterns to predict tumor type). The undergraduate researcher will explore several publicly available genomic datasets, identify and visualize the effects of these variables, apply a machine learning algorithm and assess its performance.

Supervisors: Prof. Johann Gagnon-Bartsch and Nora Payne

4. Estimating causal effects using cluster-randomized experiments: Applications in education

Randomized experiments are a common and highly regarded tool for estimating the effect of an intervention. While the randomization mechanism allows us to be less concerned about confounding from observed and unobserved factors, there may remain some imbalance in relevant characteristics between treatment and control groups even after randomization. Therefore, it can be beneficial to include information from covariates in order to improve treatment effect estimation from experiments ("covariate adjustment"). Motivated by studies of interventions in education, where randomization may be at the school or class level while the outcome of interest is at the student level, we will focus on cluster randomized and pair-cluster randomized experiments. We will compare cutting-edge methods for covariate adjustment of treatment effect estimates for cluster randomized experiments, carrying out simulations in R to obtain variance and mean squared error and investigate ways to visualize the results.

Supervisors: Prof. Johann Gagnon-Bartsch and Charlotte Mann

5. Exploration of cancer drug screening data

The student researcher will be given large and complex data from cancer drug screening experiments. The data will include information on the effectiveness of hundreds of drugs tested on hundreds of cell lines. The drug screening data contains widespread measurement error, which causes problems during analysis. With the ultimate goal of improving personalized cancer treatment, the student researcher will explore the drug screening data and adapt methods of measurement error detection. The student researcher may also develop simulations of such drug screening data to improve experimental design methods. The student researcher will learn to work with a variety of real-world, messy data and will use methods to integrate different types of complex data. Ideally, the student researcher will be comfortable coding in R.

Supervisors: Prof. Johann Gagnon-Bartsch and Zoe Rehnberg

6. Information spread on social networks

The use of social networks to spread information is ubiquitous. Influence maximization (IM) algorithms typically use social networks to select individuals who will generate the greatest spread if seeded with information. However, in social networks with community structure, most IM algorithms may yield significant disparities in information coverage between communities, which could be problematic in settings such as public service messaging. Interestingly, there are many parallels between modeling the spread of information and the spread of infectious disease. Much of the literature on disease spread uses continuous models such as differential equations to describe disease spread over time. So far, our research has focused on discrete models of information spread. This project will focus on researching continuous models of disease spread within networks with community structure and adapting them for information spread.

Supervisor: Dr. Octavio Mesner

7. Modeling and inference for cholera in Haiti

Public health decisions must be made about when and how to implement interventions to control an infectious disease epidemic. These decisions should be informed by data on the epidemic as well as current understanding about the transmission dynamics. Such decisions can be posed as statistical questions about scientifically motivated dynamic models. We will study how to ask and answer such questions in the context of a cholera epidemic in Haiti. The 2010 introduction of cholera to Haiti led to an extensive outbreak and sustained transmission, eliminated in 2019 with the help of vaccination and other public health measures. We will study four models developed by expert teams to advise on vaccination policies. We assess methods used for developing, fitting, and evaluating these models, and seek recommendations for future studies.

Supervisors: Prof. Edward Ionides and Jesse Wheeler

8. The impact of the new heart allocation policy on heart transplantation

Each year, over 7300 individuals in the United States are on waiting lists for a heart transplant. However, only about 3200 transplants are performed annually, with a 10% waiting list mortality rate between 2015 and 2017. The new heart allocation policy enacted in October 2018 was meant to standardize how available hearts are distributed with the goal of more equitable access to transplantation nationally, and it has changed the way in which hearts are allocated in the U.S. In this project, we will use data from the Scientific Registry of Transplant Recipients to assess the impact of the new heart allocation policy on: (1) the regional variation in the waiting-list and transplant outcomes; (2) the treatment practices for transplant candidates; and (3) the distributions of eligible donor characteristics. The undergraduate researcher will join a team consisting of both clinicians and statisticians to evaluate the policy effect, and the work involves (1) data integration and exploration; (2) statistical modeling and inference; and (3) software implementation in R.

Supervisors: Prof. Ji Zhu, Jinming Li, and Weijing Tang

9. Projection pursuit for dyadic data

We will develop projection pursuit algorithms for uncovering the dyadic structure in data measuring circadian patterns. We have minute-resolution data on heart rates for cancer patients and care-givers (the dyads) for several months. We are interested in extracting the variance components from these data that are stable within individuals and/or shared between members of a dyad. Students must have completed Math 214 or a more advanced linear algebra class, and completion of Stat 415 is strongly preferred. Students working on this project should be enthusiastic about devising, implementing, and evaluating algorithms.

Supervisor: Prof. Kerby Shedden

10. Diagnostic classification in educational measurement

Cognitive Diagnosis Models (CDMs) are popular statistical tools widely applied to educational assessments and psychological diagnoses, which have been receiving increasingly more attention in the past two decades. In many modern assessment situations, examiners are concerned with specific attributes that the examinees possess, and thus a simple overall score is no longer sufficient to depict the whole picture of the candidates. As a result, a finer evaluation of the examinees' attributes is desired. CDMs are such tools. They model the relationship between the test items and the examinees' latent skills, which is helpful in assessment design and post-assessment analysis of the examinees' latent attribute patterns. Estimation of CDMs with many items and latent attributes from observational data has been a big challenge due to its high computational cost. This project involves using machine learning techniques to overcome the computational difficulties.

Supervisors: Prof. Gongjun Xu and Chenchen Ma.

11. Educational fairness

To promote diverse and fair education, one fundamental issue of education testing is to ensure that the test items provide a fair comparison among different subpopulations, such as the gender and race of the students. Establishing such measurement invariance property of a questionnaire or test is a key step for establishing its measurement validity and fairness. Measurement invariance is typically assessed by differential item functioning (DIF) analysis, i.e., detecting DIF items whose response distribution depends not only on the latent trait (such as a student's ability level) measured by the test but also the group membership (such as gender or race). This project will focus on the DIF analysis in cognitive diagnosis measurement. Basic knowledge of R is required.

Supervisors: Prof. Gongjun Xu and Chengcheng Li.

12. Inference of speciation patterns from extant birth-death trees

In phylogenetics, the estimation of speciation and extinction rates from tree data is confounded by the fact that extinct branches usually cannot be observed. Recent work has focused the discussion onto so-called pulled rates that reflect the net amount of speciation over time, but inferring the underlying birth-death model remains a challenge. This presents challenges for distinct studies that assume the birth-death model can be (efficiently) inferred from trees, beyond reconstructing the Tree of Life. A couple applications are the geographic transmission of COVID-19 and the expansion of gene families during genome assembly and annotation. We aim to derive consistent estimators and their rates of convergence for old and new models as well as validate our results with simulations.

Supervisor: Dr. Brandon Legried

13. Home court advantage

In sports, many analyses have shown that there are significant home team advantages; for example, Nevo and Ritov (2013) showed that the home team in soccer receives significantly less red cards compared to the visiting team. In this project, we are interested in quantifying the difficulty of each stadium for the visiting team while controlling for various factors, such as quality of the home team. Thus, given two teams of equal quality, we want to decide which team is more difficult to play on the road. Reference: Nevo, D., & Ritov, Y. A. (2013). Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. *Journal of Quantitative Analysis in Sports*, 9(2), 165-177.

Supervisors: Prof. Ya'acov Ritov and Michael Law

14. Spatial regression discontinuity

In a recent study, Gu et al. (2017) showed that, in a given time zone, places further to the west have higher incidence rates of cancer while controlling for other variables, such as urbanization, poverty, and smoking. This suggests that sunlight in the early morning is better for our health compared to more sunlight in the evening. Compared to Gu et al. (2017), who used multiple linear regression to assess the impact on longitude within a given time zone, we are interested in using a regression discontinuity framework to analyze the effect of longitude position on other health outcomes — for example, we can consider Covid; that is, by focusing on neighboring counties in different time zones, we can directly compare their outcomes. Reference: Gu, F., Xu, S., Devesa, S. S., Zhang, F., Klerman, E. B., Graubard, B. I., & Caporaso, N. E. (2017). Longitude position in a time zone and cancer risk in the United States. *Cancer Epidemiology and Prevention Biomarkers*, 26(8), 1306-1311.

Supervisors: Prof. Ya'acov Ritov and Michael Law

15. Mitigating algorithmic biases in health risk assessment

A significant source of algorithmic bias in modern ML models is training with proxy labels. For example, in health risk assessment, a patient's healthcare cost is often used as a proxy for the patient's healthcare needs. Many proxy labels (including healthcare costs) are confounded with sensitive demographic attributes, which leads to algorithmic bias in resulting ML models. We will develop new methods for mitigating algorithmic biases in health risk assessment tools trained with healthcare cost as a proxy label for healthcare needs. The undergrad researcher will:

1. Develop statistical models which explain the algorithmic biases in healthcare risk assessment tools,
2. Design and implement algorithms to mitigate algorithmic biases in risk assessment tools,
3. Demonstrate the efficacy of the methods on a small synthetic electronic health record dataset.

Supervisors: Prof. Yuekai Sun, Prof. Moulinath Banerjee and Subha Maity

16. Dynamic network modeling of political interactions

This project will analyze publicly available data produced for the Integrated Crisis Early Warning System (ICEWS), a database of political interactions. Interactions represent directed events between two political actors, and can be used to construct networks among actors, or among their nations of origin. The summarized interaction events are indexed by time, monthly from 2005 to 2014, so the data will consist of multiple networks with a time-varying component. The student will use existing and new statistical methodology to analyze this rich dataset, fitting "dynamic" network models which summarize the networks while also sharing information across time. They will contribute to data pre-processing, fitting different statistical models (using either publicly available packages or code provided by the project supervisors), and interpretation and visualization of output.

Supervisors: Prof. Liza Levina, Prof. Ji Zhu and Peter MacDonald

17. Post-Selection Inference with Applications in Neuroscience

With complex modern data, it is common to first select a model based on an exploratory analysis. If this model is then fitted to data and inference (p-values, confidence intervals) is performed without taking into account the model selection step, it becomes unreliable; this phenomenon has played a major role in the current reproducibility crisis. The task of post-selection inference is to account for model selection in subsequent inference. In this project, we will apply post-selection inference algorithms to datasets from fMRI brain imaging, where predictors are (high-dimensional) brain connectivity matrices of subjects, along with other covariates, and responses are typically scores on cognitive or diagnostic assessments. The student(s) will apply the methods we have developed for this task to a number of brain imaging datasets, and be responsible for summarizing and visualizing the results. Familiarity with Python is required.

Supervisors: Prof. Snigdha Panigrahi, Prof. Liza Levina, Natasha Stewart