Siyi (Luke) Wei

MASTER STUDENT OF STATISTICAL SCIENCES AT UNIVERSITY OF TORONTO

☎ (647) 870-8349 | ✉ weisiyi2@gmail.com | www.wei-siyi.com | superp0tat0 | in siyi-wei

Education

University of Toronto
Toronto, ON

MASTER OF SCIENCE IN DEPARTMENT OF STATISTICAL SCIENCE (CGPA: 3.8/4.0) *Aug 2020 - Aug 2021 (Expected)*

- Activities: CUPE teaching assistant funds ($10000) and Graduate Teaching Assistant for multiple courses.

HONOURS BACHELOR OF SCIENCE (CGPA: 3.3/4.0) *August 2015 - April 2020*

- Specialist in Statistics, Statistical Machine Learning and Data Science Stream, Minor in Computer Science
- Relevant Courses: Big Data Analytics, Intro & Advanced to Machine Learning, Deep Learning, Software Engineering, Data Collection, Database, Data Structure
- Teaching Assistant for Introduction to Machine Learning, Advanced Statistical Learning, Linear Algebra, Probability, Statistics and Undergraduate probability theory.

Skills

Languages	Python(Scikit-Learn, Pandas, PyTorch, PySpark), R(Tidyverse, RShiny, mlr3, qqplot2, Rstan), SQL, Git, Scala, C#
Technologies	Spark, Hadoop, Jupyter, Amazon Web Services, TensorFlow, Linux (Ubuntu), Jenkins, Selenium
Technical Skills	Statistical Modeling, Machine Learning, Data Mining, Data Visualization, Time Series Analysis, A/B Testing, CI/CD

Experience

Google Summer of Code
Toronto, ON

DATA SCIENTIST INTERN *May 2021 - Present*

- Guided by two mentors to contribute to R community by developing the mlr3 fairness package to help mlr3 users to detect and correct the fairness problems in multiple approaches. Implement debiasing strategies as pre and post processing PipeOperators in the style of the mlr3 pipelines, which supports some bias mitigation algorithms like reweighing and equalized Odds Postprocessing.
- Create popular fairness metrics like confusion matrix, AUC and ROC curves then mitigate into other mlr3 packages. Implement visualizations and a clear API for auditing using either ggplot2 or inherit from other fairness packages following OOP principles.
- Create an introduction vignette, demos and a well documented wiki page for debiasing algorithms to showcase the new package.

University of Toronto Scarborough
Toronto, ON

RESEARCH ASSISTANT *May 2019 - April 2020*

- Built interactive Data Visualization using R Shiny to help professors detect suspicious activities from students tests. Using MOSS (anti-plagiarism system) data to locate different groups of students and report their grade distribution across semesters and assignments.
- Using logistic regression to analyze the relations between multiple factors and student's pass/fail status. Applied PCA regression to ease the collinearity founded in dataset. Design hypothesis tests to verify potential hypotheses made from exploratory data analysis.

Rakuten Kobo
Toronto, ON

QA AND DEVELOPER INTERN *January 2019 - April 2019*

- Worked on development and QA for new Kobo Audio subscription services using Jenkins, Ruby and selenium. Wrote SQL and Ruby scripts for regression testing and unit testing. Collaborate with Manager to design A/B Testing to improve current features.
- Responsible for the QA automation and bug tracking system for 4 months, prevent service failure for multiple times and received praise from team lead and promoted to work on web development and ETL pipeline for new subscription.
- Developed web page and ETL pipeline using C# and MongoDB via CI/CD and Agile development, automated the data collection process and developed new metrics for subscription data. Received praise from Big Data Team for increasing their efficiency of the data collection process.

Projects

Data Fairness correction with Bayesian Hierarchical Model

- Verify the disadvantageous position of African American in COMPAS dataset by reporting the subgroup false positive and false negative rates using classification trees. Compare the accuracy and subgroup performance of logistic regression, classification trees and the COMPAS algorithm. Identified the different subgroup patterns and imbalanced observations through Exploratory Data Analysis.
- Implement different hierarchical logistic regression models with prior information in stan to correct the algorithmic bias. Decide using a three stage bayesian hierarchical model to adjust the subgroup fairness. Successfully balanced the subgroup fairness by reducing 67%FP-37%FN to 48%FP-52%FN for African American with only 1.7% loss in total accuracy. Further extend this project to GSoC 2021.

Improved Monte Carlo Tree Search (Reinforcement Learning project)

- Implement the naive Monte Carlo Tree Search (MCTS) algorithm in Python with board games. Then improve the simulation policy of naive MCTS and make it workable on larger game boards. Which could balance the trade off between winning rate and time complexity in fewer simulations. Achieve a 75% winning rate compare to naive MCTS with fewer simulations.

Unsupervised pre-training with spectral features

- Conduct research with professor and raise hypotheses on spectral features that could be learned by convolutional layers. Working on empirical study with Google Colab, pytorch, pandas and other machine learning libraries. Extract spectral features using different approaches like Kernel PCA or Convolutional layers. Those spectral features could be useful in few shot learning or transfer learning.