Yilin Zhu

yz3323@columbia.edu | (858) 539-3692 | https://ericzyl.github.io/

EDUCATION

Columbia University, New York City September 2023 – December 2024
M.A. in Statistics

- Relevant Coursework: Machine Learning, Optimal Transport, Modern Analysis, Deep Learning, Computational Statistics, Inference, Regressions, Bayesian Statistics, Statistical Analysis of Neural Data, Databases

University of California, San Diego (GPA 3.7/4.0) September 2019 – June 2023
B.S. in Applied Mathematics, Minor in Computer Science

- Relevant Coursework: Probability Theory, Optimization, Graph Theory, Time Series, Combinatorics, Stochastic Processes, Data Structures, Algorithms, Computation Theory, Computer Organization

SKILLS

Software: Python, Java, C/ C++, SQL, R Programming, MATLAB, PyTorch, Stan, HTML/CSS

RESEARCH EXPERIENCE

Columbia University | Statistics Department March 2024 – Present
Advisor: **Prof. Parijat Dube** New York

Enhance Document Similarity Measures via Optimal Transport

- Implemented optimal topic transport model with linear optimal transport for document classification, achieving a 100x improvement in computational efficiency while maintaining test accuracy.
- Built document similarity metrics with SBERT, optimizing concept-based queries to help users retrieve specific cases in legal documents quickly and accurately.
- Implemented an unsupervised sentence ranking method in long articles using KSVD and PageRank, combining intra-article and corpus-wide relevance to enhance document summarization tasks.
- Explored graph optimal transport for document similarity.

Develop Language Model for Improved Topic Analysis in Police Narratives

- Developed Python pipelines and tuned BERT with UMAP for better topic representation in police narratives, aiding decision analysis and optimizing report accuracy for law enforcement.
- Visualized criminal patterns over time with interactive temporal charts, highlighting shifting themes for trend analysis, and presented results at an industry event to discuss with professionals.

Columbia University | DitecT Lab June – August 2024
Advisor: **Prof. Sharon Di** New York

Build Robust & Explainable GNN for Spatial-temporal Predictions

- Implemented multimodal mobility nowcasting using EAST-Net and ST-Net models to predict mobility patterns, utilizing datasets representing urban mobility trends in major cities and regional trends during the pandemic.
- Managed data input, normalization, and splitting into training, validation, and test sets for large-scale mobility datasets, configured and executed training scripts for both models using PyTorch, and evaluated model performance with RMSE, MAE, and MAPE metrics.
- Enhanced models with temporal covariates, integrated Heterogeneous Mobility Information Network (HMIN), and applied Memory-Augmented Dynamic Filter Generator (MDFG) for dynamic parameter adjustment.
- Analyzed SHAP values for feature importance and conducted evaluations via t-tests.

Legal Aid Society | DNA Unit January – July 2024
Advisor: **Prof. Daniel Rabinowitz** New York

Access Algorithmic Fairness in Forensic Probabilistic Models

- Designed algorithms to conduct goodness-of-fit test using Wasserstein Distance to evaluate the null hypothesis of assumed distributions in forensic models, ensuring algorithmic fairness.
- Analyzed statistical assumptions in forensic models and designed hypothesis tests, equipping attorneys with data-driven arguments, strengthening case strategies and outcomes in juvenile trials.
- Investigated forensic algorithmic integrity, particularly Metropolis-Hastings algorithm convergence.

University of California-San Diego | Finance Department March 2022 – February 2023
Advisor: **Prof. William Mullins** San Diego

Analyze Celebrities' Influence on Crypto Markets

- Built crypto promotion databases using Twitter and TikTok APIs, identifying positive buy signals via RoBERTa.
- Conducted regression analysis to access the effects of social media influencers on financial markets.
- Implemented web automation to create partially identified SSN database, ensuring privacy while identifying immigrants for research studies.

PROJECTS

Data Integration via Gromov-Wasserstein October 2024 – Present

- Aligned cross subjects neural spike data with fused Gromov Wasserstein distance and incorporated linear optimal transport for improved computation efficiency, enhancing prediction of monkey behavior in center-out tasks,

Wine Quality Analysis May 2024

- Implemented regressions, ANOVA, and MCMC to assess the impact of physicochemical properties on wine quality, addressing data contamination for accurate statistical analysis.
- Optimized Gibbs Sampler by selectively updating covariance, maintaining data imputation accuracy, and reducing runtime, accelerating project experiments and timely analysis completion.

Optimal Transport for Image Color Processing February 2024

- Implemented Sinkhorn-Knopp algorithm with Python, solving the entropic regularized OT in color transferring.
- Trained a KNN model to map source image colors to the target's distribution, reconstructing and converting the transformed image for visualization.

Recipes Website Database Application October – December 2023

- Designed E/R diagrams and constructed a PostgreSQL database for a recipe website.
- Engineered a recipe-sharing web platform leveraging Python Flask and SQLAlchemy, integrating functions like user authentication, recipe look-up, uploads, saves, and reviews.
- Created a dynamic user interface with HTML and JavaScript, deploying the application via Google Cloud Platform.
- Implemented a collaborative filtering recommendation system to provide personalized recipes suggestions to users.

Telecom Customer Churn Prediction May – June 2023

- Applied forward feature selection with AIC, designed exploratory data analysis using group bar chart.
- Utilized machine learning methods such as XGBoost, SVM, Regressions, and Random Forest, contributing to accurate customer churn forecasts, and empowered proactive decision-making.
- Evaluated ML models via cross validation to ensure a robust model, achieving AUC score of 0.92.

HIV Transmission Graph Application March 2023

- Developed a C++ Graph generator in C++ to analyze HIV data. This tool read input edge list from CSV files and facilitates essential graph operations including neighbor and edge weight retrieval.
- Implemented Dijkstra's Algorithm and Up-Trees data structure to find weighted shortest paths, connected components, and smallest connecting threshold, identifying transmission clusters and infection pathways.
- Created Huffman Coding Tree to compress files, optimizing the storage of large HIV datasets

PRESENTATION

- "Linear Optimal Topic Transport for Document Similarity" Presented at UC Davis Peter Hall Conference on Statistics in the Age of AI 2024
- "Alternating Direction Method of Multipliers with Applications" Presented at Computational Statistics Seminar at Columbia University 2024.
- "Developing Language Model for Improved Topic Analysis in Police Narratives" Presented on Data Science Day at Columbia University 2024.

HONORS

- Provost Honors 2019 – 2023

VOLUNTEER

- Notes taker in Applied Linear Algebra and Combinatorics for students with disabilities.