Shiyu Ma

sylviama1026@g.ucla.edu | sylviama1026.github.io/

Education

UNIVERSITY OF CALIFORNIA, LOS ANGELES (UCLA) Los Angeles, CA
B.S. in Mathematics / Economics and Statistics, Specialization in Computing Expected June 2023
Cumulative GPA: **3.97/4.0**, Deans Honors List(All Quarters)
Relevant Coursework: Real Analysis, Stochastic Processes, Linear Algebra, Optimization, Numerical Methods, Probability and Statistical Inference, Statistical Models and Data Mining, Data Analysis and Regression, Linear Models, Computational Statistics with R, Python with Applications, Monte Carlo Algorithms, Experimental Design

Software

Ma, S., Zou, L. (2022). **scGTM**: Single-cell generalized trend model(R package: github.com/Sylviama1026/scGTM)

Research Interest

Statistical Machine Learning, Applications in Genomics and Bioinformatics, Model Interpretability and Robustness

Research Experience

Junction of Statistics and Biology Lab Los Angeles, CA
Research Assistant, Advisor: Prof. Jessica Jingyi Li Mar. 2022 - Present

scGTM: Single-cell generalized trend model

- Designed R package for scGTM(Cui et al., 2022) that uses GAM motivated model to fit gene expression trends along cell pseudotime that can both flexibly capture gene trends and generate biological interpretable parameters
- Customized initial condition designs in the Particle Swarm Optimization (PSO) to stably estimate different sets of parameters in Poisson, Negative Binomial, ZIP, and ZINB marginal distributions of gene expression counts
- Derived MLE scheme and confidence interval for Gaussian marginal distribution to enlarge cost function choices
- Modified fitted curve formula of the original model to enable removal of potential confounding factors
- Developed annotated visualizations to highlight the potential impact of zero inflation on the fitting trend
- Compared scGTM with GAM, GLM, LOESS, switchDE, and ImpulseDE2 on real single cell RNAseq datasets
- Showed empirical coverage probability of parameters is consistent with theoretical confidence interval derived from partial Fisher information matrix

ITCATree: Hierarchical cell types tree generator based on Information-theoretic Classification Accuracy

- Built ML-based cell type similarity tree to refine ambiguity and subjectivity in cell type annotation
- Derived non-Euclidean cell types distance metric from label combination order suggested by entropy-weighted classification metric(Zhang et al., 2022) that can balance accuracy and classification resolution
- Designed an algorithm to assign consistent relative distance between cell types in growing combined groups
- Compared the generated tree across classifiers KNN, LDA, Random Forest, and SVM using 4 scRNAseq datasets, identified the effectiveness of weak classifiers in suggesting label combination
- Experimented various PCA parameters settings to show the generated cell types trees are more robust to data preprocessing procedure than hierarchical clustering due to machine learning model's resistence to noises

PARIS Lab Los Angeles, CA
Research Assistant, Advisor: Prof. Mathieu Bauchy Feb. 2022 - Present

Predicting Fracture Energy of Porous Materials by Symbolic Regression

- Optimized feature selection and design in Symbolic Regression to predict material's fracture energy with interpretable geometric features, achieved accuracy comparable to the CNN and improve interpretability
- Investigated algorithm implementation differences between Matlab and Python packages to guide parameter tuning
- Standardized and transformed raw data based on each distributions to reduce the solution search space of the algorithm, the running time, and the complexity of results
- Conducted SHAP analysis to identify error sources and correct feature bias, increased prediction accuracy by 20%
- Quantified interaction effect between geometric features using Hessian matrix to reveal rules of material properties
- Implemented simulated annealing and steepest descent algorithms to generate high fracture energy material designs

Work Experience

Thumbtack San Francisco, CA
Data Analyst Intern Jun. 2022 - Sep. 2022

- Built Random Forest and Logistic Model end-to-end to predict the sales conversion probability of potential users
- Solved imbalance data issue using stratified sampling and cost-sensitive learning on the two models respectively
- Led feature engineering to remove multicollinearity and cut model features from 52 to 5 with 5% increase in recall
- Developed function based on random forest to generate metrics and features' ranks from multiple rounds of sampling to effectively eliminates model differences due to randomness and suggest feature selection
- Designed metrics, Tableau dashboard, and power analysis to verify 8% revenue increases in marketing AB testing

TAL Education Group Beijing, China
Data Scientist Intern Jun. 2021 - Sep. 2021

- Conducted Clickstream Clustering to identify user subgroups' needs by Python, raised conversion rate by 5%
- Designed revenue, cost, and satisfaction metrics using SQL and Excel to optimize product lines, cut costs by 8%
- Drove model-based strategy cross-functionally with the operations team to improve user expereinces

Onomy Los Angeles, CA
Data Scientist Intern Mar. 2021 - Jun. 2021

- Scraped 47,812 Reddit posts for text mining to understand adulting concerns and related business chances
- Performed NMF topic modeling and emotion decomposition for 8 keywords to identify user subgroups' needs, offered suggestions of marketing tone and strategies in a 20-pages report

Projects

Stochastic SIR Model on SARS-CoV-2 Omicron transmission in China May. 2022 - Jun. 2022

- Derived theoretical proof to compute expected ending time and ending probability of local Omicron transmission
- Generated simulation for infection process as asymmetric random walk with validated real parameters

GAN(Generative Adversarial Network) Apr. 2021 - Jun. 2021

- Trained DCGAN model using TensorFlow on Cats faces dataset, evaluated by FID score with Python
- Demonstrated correlation between FID score and image data by implementing increasing noise levels of images

Price Prediction Model for Uber and Lyft Oct. 2020 - Feb. 2021

- Built price prediction models for Uber and Lyft using LASSO, Random Forest, and Gradient Boosting models
- Created heatmaps in R to identify price changes over locations and rush hours with cost-saving recommendations

Skills

Python(Pandas, NumPy, Matplotlib, Scikit Learn, Plotly, TensorFlow, Scipy), R, SQL, C++, Linux, Git, Matlab

Honors & Rewards

Queen's Road Foundation Undergraduate Research Fellowship, *1 Recipient per Department, 2022*

ASA DataFest at UCLA, *Finalist, 2022*

Danaher Scholarship*, 2019-2022*

Leadership Experience

Rye International, *Head of Data Science Track* March. 2021 - Jun. 2022

- Initiated weekly study groups and panels in statistical machine learning among international students
- Facilitated data science project collaboration with a local start-up, published final reports as 3 articles on Medium
- Mentored junior students in academics, course planning, campus life, and career development

Chinese Students and Scholars Association, *VP of Career Development* Sep. 2019 - Jun. 2022

- Led a team of 40 to organize Workshops and Panels series with 500+ participants, raised more than $5000 a year
- Initiated and operated online social groups with 600+ participants for sharing internship/research opportunities
- Hosted online recruiting events across 3 campuses for top companies such as Tencent, Alibaba, and ByteDance