KAI LIU

kai.liu@utexas.edu • 512.917.6781
GitHub: github.com/KaiUT • LinkedIn: www.linkedin.com/in/kai-liu-utaustin

PROFILE

Ph.D. in Computational Biology (expected May 2019) seeking internship opportunities in Data Science/Machine Learning Research for 2018 summer. Solid skills in machine learning, statistical inference. Comprehensive computing skills including Python, R, SQL, Hadoop, Spark, TensorFlow. Excellent problem solving skills in both independent and team environments. Skilled presenter of technical materials to both technical and non-technical audiences. Quick, thorough and effective learner.

EDUCATION

Ph.D. in Computational Biology ◇ *The University of Texas at Austin, Austin, TX* — Expected May 2019
M.S. in Microbiology ◇ *Huazhong Agricultural University, Wuhan, China* — Grad. June 2013
B.S. in Biotechnology ◇ *Huazhong Agricultural University, Wuhan, China* — Grad. June 2011

WORK EXPERIENCE

Machine Learning Intern ◇ *QuintilesIMS, Plymouth Meeting, PA* — June 2017 - August 2017

Predicted Quality of Investigators in Future Clinical Trials | Python, Spark

- Predicted outliers of investigators per Key Risk Indicator using distribution based approach;
- Built multiple machine learning models (Lasso Regression, Neural Network, Random Forests) to predict the quality of investigators in a future study, which is one of the core projects in the investigator recommender system.

Graduate Research Assistant ◇ *The University of Texas at Austin, Austin, TX* — December 2014 - Present

Developed Infectious Diseases Surveillance App | Python

- Developed a regression model and a Multivariate Exponentially Weighted Moving Average (MEWMA) model to detect emerging outbreaks with an accuracy of 0.9;
- Combined above models with stepwise variable selection algorithms to select best data sources for infectious diseases surveillance (more than 400 data sources in total);
- Built up data pipeline to automate the process of retrieving and cleaning data from CDC, RSS feed, Google Trends, Wikipedia, Twitter etc; integrated the App into Cloud Ecosystem.

PERSONAL PROJECTS

Being Involved in Building an Open Source Software to Detect Lung Cancer | Python, TensorFlow August 2017 - Present

- Contributing to improve the 3D Convolutional Neural Network that identifies locations of nodules in scans;
- Contributing to improve the 3D Convolutional Neural Network to find the boundaries of nodules in scans.

Developing a R Package for Big Data Analysis | R & Rcpp — December 2016 - Present

- Implementing following algorithms in the package: Stochastic gradient descent using line search and quasi-Newton methods to determine step size · The lasso · The proximal gradient method · Laplacian smoothing solved by sparse Cholesky/LU, the Gauss-Seidel method, the Jacobi iterative method, and conjugate gradient method · Graph fused lasso solved by Alternating Direction Method of Multipliers (ADMM) · Sparse matrix factorization.

Predicted the Direction of Exchange-Traded Fund (ETF) movement | Python — April 2017 - May 2017

- Retrieved nine histrorical ETF sector datasets from Yahoo Finance;
- Implemented Logistic regression, Ridge & Lasso regression, and Artificial Neural Network to predict the direction of ETF movement;
- Achieved an accuracy of 55% $\sim$ 60% for predicting nine ETF sectors movement; and the trading strategy based on my prediction outperforms baseline strategies.

Predicted Yelp Rating Based on User Review Enhanced Collaborative Filtering | R — September 2015 - December 2015

- Extracted user opinions from restaurants dataset from Yelp ($\sim$10GB) using Stanford coreNLP tool;
- Developed a new Collaborative Filtering-based method to improve the accuracy of user's rating prediction and solve the sparseness of dataset by combining item's features and user opinions from all reviews;
- Improved the prediction accuracy by 4.23% compared to the traditional KNN method, and the coverage is 100%.

SKILLS

Programming	Fluency in Python(NumPy, SciPy, pandas, scikit-learn), R, Git · Familiar with MATLAB, Linux, LaTex · Experience in SQL, Hadoop, Spark, TensorFlow, C++
Machine Learning	Deep Neural Network · Regression with regularization · Support Vector Machine · Random Forests · Hidden Markov Model · Clustering · Time series and dynamic models · Frequent Pattern Mining · Natural Language Processing · Image Processing