jxx151730@utdallas.edu Cell: (469) 835-4885

Jing Xia

Address: Dallas, Texas 75082

Data Scientist/Bioinformatics Scientist

Summary

PhD candidate in bioinformatics, AI enthusiast who are interested in applying novel Machine Learning methods to solve biological questions. Strong statistical background and solid programming skills in Python, SQL and R.

Education

University of Texas at Dallas *Richardson, Texas* Expected December 2022
Ph.D., Bioinformatics GPA: 3.74/4.0
University of Texas at Dallas *Richardson, Texas* Expected December 2022
Master of Science in Data Science GPA: 3.84/4.0
Coursework: Machine Learning, Advanced Statistical Methods (I&II), Multivariate Analysis, Statistical Inference, Big Data, Graph Theory, Data Structure & Algorithm, Quantitative Biology

SKILLS

Programming: Python, R, SAS, Bash, SQL, Apache Spark, PyTorch, TensorFlow, Scikit-learn, Advanced Excel
Machine Learning Techniques:

- Linear Regression
- Logistic Regression, Linear/Quadratic Discriminant Analysis, KNN, Random Forest, SVM
- Neural Network, RNN(LSTM), CNN
- K-Means, Hierarchical, Latent Dirichlet Allocation
- Principal Component Analysis

Career Experience

University of Texas at Dallas *Richardson, Texas* January 2018 – December 2022

PhD Projects - Bioinformatics

RNA Variants Calling by Rigorous Statistical Framework

- Built a dictionary-based annotation system in Python to map billions of RNA sequencing sites to the genome.
- Designed and implemented pipeline on clustering-based computing system to analyze 426 cancer cell RNA sequencing data by comparing three RNA identification packages.
- Applied Binomial Test and Chi-squared Test on ~4 millions of RNA variants site to filter out systematic errors and obtained ~2 millions of RNA editing sites

Causality Inference between RNA Editing and RNA Binding Protein

- Identified ~1 million common RNA editing sites from multiple datasets.
- Conducted Fisher's Exact Test to find biological significant editing sites. Followed by several multiple test correction methods (Bonferroni and FDR).

Protein Binding Prediction with RNA Sequence Data

- Preprocessed sequencing data and encoded sequence in one-hot encoding in PyTorch.
- Trained a Convolution Neural Network to predict the binding of a specific RNA binding protein.

Master Projects – Data Science

Comprehensive linear Regression Analysis of 1000 Cardiovascular Patients' Data in R

- Performed linear assumption check by Shapiro-Wilk Test, Brown-Forsythe Test and Lack of Fit Test. Outliers were check by Studentized Residuals.
- Carried out Box-Cox transformation to get normality of data.
- Detected multicollinearity by Variance Inflation Factor and Condition Index.
- Implemented three forms of linear models: Ordinary Least Square, Lasso Regression and Ridge Regression. Hyper parameters were optimized by Leave-One-Out Cross Validation.
- Estimated test MSE for each model by k-fold cross-validation.

In-depth Classification Analysis of 569 Breast Cancer Patients in Python

- Preprocessed data by data cleaning, categorical feature transformation and standardization, etc.
- Trained supervised machine learning models including Logistic Regression, KNN, LDA, QDA, SVM, Random Forest and Neural network. Regularization was applied to avoid overfitting with parameters determined by k-fold cross validation. Optimization methods such as Adam and mini-batch gradient descent were applied for Neural Network to speed up training. Various kernel functions were utilized for SVM.
- Coded numerical searching algorithm (Newton-Raphson) to solve the likelihood function of logistic regression.
- Evaluated performance (Accuracy or AUC) of each classifier via k-fold cross-validation and obtained the feature importance to find top factors for the prediction.

Stock Price Prediction based on Deep Learning in Python

- Exploratory data analysis of 15000 S&P 500 records followed by data normalization and splitting.
- Built a simple RNN model from scratch to predict the open price. Truncated Backpropagation Through Time algorithm was coded to save computational power and regularization terms were implemented to avoid gradient explosion. Fine-tuned the model by trying different combinations of hyperparameters.
- Trained LSTM and GRU models via TensorFlow on GPU to validate the performance of hand built RNN. Various activation functions were tested.
- GRU outperformed naïve RNN and LSTM on the testing data with respect to Mean Square Error.

Natural Language Processing and Topic Modeling of 5000 Cell Phone Reviews in Python

- Pretreated customer reviews by tokenization, stemming, removing stop words. Extracted features by Term Frequency – Inverse Document Frequency (TF-IDF).
- Utilized three unsupervised learning models of K-means Clustering, Hierarchical Clustering and Latent Dirichlet Allocation.
- Recognized latent topics and key words of each review for clustering.
- Visualized training results by Principal Component Analysis (PCA).
- Three algorithms provided meaningful clustering of customer reviews with latent semantic structures retrieved.

UT Southwestern Medical Center *Dallas, Texas* January 2018 – May 2019

Research Assistant – Biochemistry & Bioinformatics

Participated in research project on biomedical data, applying molecular biology techniques and t-test for the biological difference.

- Publication: Ebrahim H.GHazvini Zadeh, ZhiJiang Huang, Jing Xia, Dalliang Li, Howard W. Davidson, and Wen-hong Li, Functional study of ZnT8 in pancreatic cells by immunofluorescence and electron microscopy and quantified the expression level of three different pancreatic hormones in hundreds of pancreatic cells. (2020). https://doi.org/10.1016/j.celrep.2020.107904