Jianing (Julia) Chen

jchen801@usc.edu · LinkedIn · (765)337-0250

Objective

I am an energetic and passionate master Student working towards a MS in Applied Data Science. I am aiming to use my knowledge of statistics, database management, machine learning, deep learning to satisfy the data science/analytic internship at your company.

Education

University of Southern California, Los Angeles, CA Jan. 2021 – Dec. 2022

Master of Science in *Applied Data Science*, *GPA*: 3.9/4.0.
Coursework: Machine Learning, Database Management, Natural Language Processing, Data Mining

University of California, Santa Barbara (UCSB), CA Sep. 2019 – Dec. 2020

Bachelor of Science in *Statistics and Data Science*, *GPA*: 3.8/4.0.
Coursework: Bayesian Analysis, Regression Analysis, Time Series, Survival Analysis, Optimization

Academic Projects

Image Processing - Lung Image Classification and Diagnose Apr. 2021 – May. 2021

- Worked on over 10k+ 64*64 pixels lung images and utilized convolutional neural networks using **Tensorflow**.
- Conducted data preprocessing, including resizing data, and converting to NumPy arrays.
- Trained LeNet-5 and VGG-16 architecture; fine-tuned hyper-parameters by changing filter size and learning rate; performed data augmentation to lower effect from imbalanced data.
- Designed an ensemble of 8 models to take trade-off between precision and recall into consideration, thus final model reached classification precision of 92.12% and recall of 92.57%.

Unsupervised Learning Project - Clustering on Genetic Fingerprint Mar. 2021

- Analyzed 20,000+ patient fingerprints and intended to distinguish people who share same immunity.
- Applied PCA to reduce redundant variables from 512 dimensions to 7 dimensions.
- Performed K-means clustering and determined number of clusters using Elbow Plot and Silhouette Score.

Supervised Learning Projects - Classification for Treatment Feb. 2021

- Extracted insights by visualizing relationships among predictive features; utilized hypothesis testing and backward selection to define variables; implemented data transformation to amplify significance of statistic.
- Tested multiple classification models such as Random Forest, Logistic Regression, and Gradient Boosting Machine; performed hyper-parameter tuning with GridSearch cross-validation to minimize RMSE.
- Selected final model base on ROC-AUC score, the best model gives accuracy of 82.47%.

Data Management - Track COVID-19 in California Hospitals Jan. 2021 – May. 2021

- Worked in group of two, responsible for implementing database query, filter, and sort functions.
- Retrieved Hospital data from API save to **MySQL**; fetched streaming COVID-19 data and store in **MongoDB**.
- Implemented parallel processing to sum over each feature using **Spark RDD**.
- Built **React** and **Streamlit** web layout and optimized site design to increase user engagement.

Skills

- Programming Language: Python, SQL, R, SAS, C, Java, LaTex
- Database and Cloud: MySQL, MongoDB, AWS EC2, Firebase, Hadoop Spark, Hadoop MapReduce
- Framework: Altair, StatsModels, Scikit-Learn, Numpy, Pandas, TensorFlow, PyTorch
- BI Tools: MS Excel, Tableau Desktop, Tableau Server