Alice Wu Lim, PhD

Remote, any time zone
Los Angeles, CA 90048
(862) 485 - 3465
limwualice@gmail.com
http://www.alicewulim.com

Experience

Data Scientist
City of Hope Medical Center

Used an unsupervised learning model and advanced data analysis to discover statistically significant correlation between breast cancer genomic clusters and tnm-staging as well as breast cancer biomarker sets.

How this was accomplished:

- Formulated the problem: Set goals of (a) using unsupervised learning model to cluster patient genomic data, (b) finding statistical correlation between engineered patient features and the detected clusters
- Collected big data: Queried terabytes of data from over 30 satellite hospital databases; transformed and combined data into cleansed dataframes with hundreds of features
- Clustered data: Constructed k-modes clustering model and detected genomic clusters of patient data
- Engineered features: Used domain knowledge to engineer dozens of new features from patient data to test for correlation with genomic clusters
- Performed statistical analyses: Performed non-parametric significance testing which revealed statistically significant correlations (p less than 0.01) between several patient features and the genomic clusters
- Communicated findings: Presented the findings to business partners, collaborators, and non-experts

Tech used:

- SQL, including window functions, user-defined functions;
- AWS (Poseidon, DNAnexus)
- Pyspark
- Python, including pandas, scikit-learn, numpy, plotly (e.g., violin plots, raincloud plot, histograms)
- Algorithms: t-SNE, Kneedle, k-modes clustering
- Command line

Data scientist
Freelance work

Trained machine learning models to make predictions for real-world problems.

How this was accomplished:

- Consulted with experts: reached out to experts to obtain relevant domain knowledge for each problem
- Cleansed the data: extensive data preprocessing, including significant data imputation, feature selection techniques, and feature engineering
- Trained models: Trained several candidate models using cross-validation; tuned hyperparameters; used relevant metrics to evaluate models

Tech used:

- *Python*, including Pandas, Numpy, Scikit-learn, Matplotlib, Plotly;
- *Modeling algorithms*, including linear regression, logistic regression, random forest, XGBoost;
- *Pyspark*;
- *SQL*, including window functions and common table expressions

Education

PhD, Mathematics	MS, Mathematics	BS, Mathematics
Syracuse University	Syracuse University	UCLA

Research papers

1. Clustering whole-exome sequences of breast cancer reveals association with staging and molecular subtype, in preparation
2. The Splitting Theorem and Topology of Noncompact Spaces with Nonnegative N-Bakry Emery Ricci Curvature; Proceedings of the AMS, May 2021; https://doi.org/10.1090/proc/15240
3. Locally Homogeneous Non-gradient Quasi Einstein 3-Manifolds; Advances in Geometry, January 2022; https://doi.org/10.1515/advgeom-2021-0036