Cheng Ju

Mobile: (1) 510-316-0187, *E-mail:* jucheng1992@gmail.com
Github: https://github.com/jucheng1992 Home Page: https://jucheng1992.github.io/

SUMMARY

I am a second year PhD student in Biostatistics. My research interests mainly focus on machine learning and causal inference. My skill includes:
Proficient: R, Machine Learning, Causal Inference, Spark (pySpark)
Familiar: Python (numpy, scipy, pandas, scikit-learn, pySpark), SQL, Hadoop

EDUCATION

University of California, Berkeley, Berkeley, CA **August 2014 - May 2018 (expected)**
PhD in Biostatistics
Statistical courses: Theoretical Statistics, Probability Theory, Applied Statistics, Statistical Computing, Advanced Learning Theory
Programming courses: Data Structure, Efficient Algorithms and Intractable Problems
MOOC (find my certificates on my LinkedIn):
Introduction to Apache Spark; Scalable Machine Learning; Intro to Hadoop and MapReduce; Intro to Data Science
Southeast University, Nanjing, China **August 2010 - Jun 2014**
Bachelor in Statistics
Related courses: Multivariate Statistical Analysis, Sampling Survey, Programming and Algorithmic Language

RELATED EXPERIENCE

Machine Learning and Ensemble Learning Method for Prediction of Propensity Score **April 2015 - Present**
Research Assistant, University of California, Berkeley and Brigham and Women's Hospital
The goal of this project is to improve the estimation of propensity score using machine learning methods.

- Compare machine learning algorithms (from *caret* library in R) and ensemble learning method for prediction of propensity score based on AUC and likelihood.
- Use high-dimensional propensity score method to generate new features from health claims data.
- Apply ensemble learning method and increase AUC from 0.79 to 0.83

Collaborative Targeted Maximum Likelihood Estimation for Large Scale and High Dimensional Data **April 2015 - Present**

In this project we compare several causal models and develop variable selection method.

- Propose several pre-ordering method for variable selection of collaborative targeted maximum likelihood estimation (C-TMLE). Design different simulation cases to assess the performance of the pre-ordering methods.
- Study the performance of double robust estimators (A-IPTW, TMLE and C-TMLE) for correcting the bias of the estimation of target parameter in observation study using Monte Carlo simulation.
- This work was presented in 2015 Joint Statistical Meetings: Collaborative Targeted Maximum Likelihood Estimator (C-TMLE) in Observational Studies.

Statistical Consultant, D-lab at University of California, Berkeley **January 2015 - August 2015**

- Offer consulting service for statistical and machine learning problems. Including descriptive statistics, traditional statistical model (classical hypothesis tests, linear model) and machine learning (SVM, boosting, bagging, random forest).
- Previous projects including: PCA and visualization of high dimensional data; Penalized regression method (ridge regression and LASSO); Machine learning and model selection for prediction.

Organ Classification for Drosophila Embryo Image based on Machine Learning Method **October 2015 - November 2015**

This project mainly focuses on distinguish gut, yolk, and epidermal/mesodermal tissue in embryo images.

- Construct features from the raw embryo image based on computer vision method (e.g. Histogram oriented gradient feature and Local Binary Pattern) using *scikit-image* in python.
- Select features based on variable importance from Random Forest and Tree Boosting.
- Apply different machine learning algorithms (Random Forest, Tree boosting, Lasso, SVM) to classify the organ.
- Use cross-validation and Estimation Stability with Cross Validation (ESCV) to assess and select the models.

COMMUNICATION WORK EXPERIENCES

Graduate Student Instructor (Teaching Assistant), University of California, Berkeley **August 2014 - May 2015**

- Teach the discussion class "Introduction to Probability and Statistics" and "Big Data in Public Health".
- Prepare materials for weekly discussion. Hold office hours to help students fully understand class materials.
- Give lecture about parallel programing in R using *doParallel* and *foreach* package.

SELECTED HONORS AND AWARDS

- **The Taylor Award Scholarship,** University of California, Berkeley, 2015.