Rajan Kapoor

rkapr.github.io
r.kapoor@tamu.edu
linkedin.com/in/rkapr
☎ 979-398-2987
github.com/rkapr

Objective: Looking for internship/co-op in data science/bioinformatics.

Education:

PhD, Electrical and Computer Engineering (GPA: 3.82) May 2017 – Dec 2021
Texas A&M University, College Station, TX
MS, Electrical and Computer Engineering (GPA: 3.75) Aug 2015 – May 2017
Texas A&M University, College Station, TX
BTech, Electrical Engineering (GPA: 8.57/10) Aug 2010 – May 2014
Indian Institute of Technology, Patna, India

Data Science Projects / Experience:

***Zeroinfl*: Zero Inflated Poisson regression (ZIP) in Python** (with Eric Chuu) [Spring 2018]

- Modeled a process switching between perfect state with no errors and imperfect one with Poisson distributed errors using ZIP regression based on pscl package in R
- Implemented logit, probit, complementary log log, Cauchy and log link functions for GLM fit using OOP
- Wrote functions for likelihood, log-likelihood, gradient and max likelihood estimation using BFGS optimization, with expectation maximization-based initialization
- Wrote functions summary for pretty-printing results including p-value, standard error, z-statistics and predict for using fitted model for prediction
- Extended the functionality by including zero inflated negative binomial and zero inflated geometric regressions
- ***Key achievement:*** First open source contribution to provide zero inflated regression functionality in python

Adaboost based face detection using Voila Jones framework [Fall 2019]

- Implemented five stage Adaboost classifier in python with decision stump as weak learner for detecting face images using vertical, horizontal and diagonal Haar features
- Vectorized code to speed-up the repeated processing, observed effects of FP/FN penalty on empirical error
- ***Key achievement:*** Achieved **3.5x improvement** in training time at the cost of 2.25% reduction in empirical accuracy by filtering robust features

Minimum description length based Boolean network learning [Spring 2020]

- Implemented MDL based gene network learning algorithm from short Boolean time series by minimizing sum of error and model-description entropies
- Tested and verified the code on Boolean time series data generated using BoolNet package in R
- ***Key achievement:*** The MATLAB code based on Dougherty et al. paper was released for open source use

Gaussian Mixture Modeling for Cancer Heterogeneity [Fall 2016]

- Modeled cancer heterogeneity as mixture of Boolean networks using gaussian mixture models
- Estimated proportion of each subpopulation using expectation maximization with k-means initialization
- ***Key achievement:*** Results were published in IEEE/ACM TCBB journal

Mixture of Poissons for modeling number of daily deaths [Fall 2019]

- Derived, implemented and compared convergence of gradient descent and Newton Raphson optimization for ML estimates
- Calculated *sympy* expressions for gradient and hessian in python, then converted them to *numpy* functions

Predicting consumer behavior from shopping cart data [Fall 2019]

- Designed and implemented an efficient SVM based algorithm using one-hot coding to learn a list of items frequently bought by a predicted subclass of customers from shopping cart data

Research:

Understanding nitrogen transport pathways related to grain protein content in wheat [Spring 2019]

- Collected, cleaned, organized large alignment files from public databases using globus-cli (shell)
- Performed WGCNA based gene clustering using R scripts on computing cluster
- ***Key achievement:*** The analysis discovered two modules involved in activating storage molecules including starch, lipids and protein, discovered multiple TFs as possible key candidate genes

Gaussian Graphical model with fused lasso penalty for learning causal transcriptional regulations of Sorgoleone biosynthesis genes in Sorghum [Fall 2020]

- Inferred causal gene interactions for sorghum circadian genes using unsupervised learning for verification
- ***Key achievement:*** Identified 21 potential regulators of sorgoleone biosynthesis genes, verified TF-gene interactions using TF motifs from plant transcription factor database and comparative genomics

Skills:

- R, Python (numpy, pandas, scikit-learn, scipy, cvtools), shell scripting (SLURM, LSF), MATLAB, SQL, Jupyter/R Notebooks, C/C++
- ML models: SVM with kernels, k-means, decision trees, Adaboost, linear, logistic, generalized linear regression, parsimonious model selection (AIC/BIC), lasso, ridge regression, mixture models, mixed models, stochastic gradient, coordinate descent, random forests, time series analysis (ARMA/ARIMA/ARCH/GARCH), ANOVA.

Relevant Courses/Certifications:

- *Engineering/Statistics:* Regression Analysis (A), Statistical computing in R & Python (A), Applied Statistics & Data Analysis (B), Pattern Recognition (A), Distribution Theory (A), Information Theory (A)
- *Biology/Bioinformatics:* Bioinformatics (S), Bioinformatics Command Line (A)
- *Online Certifications:* Introduction to SQL, Learn the Command Line, Algorithms & Data Structures

Publications:

- A Gaussian Mixture-Model Exploiting Pathway Knowledge for Dissecting Cancer Heterogeneity, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018

Honors and Awards:

- Texas Engineering Experiment Station (TEES) Research Assistantship, Sept 2015-current
- Texas International Student Scholarship, Fall 2015, Spring 2016
- Runner up Best BTech Project Award in Electrical Engineering, 2014