Rajan Kapoor

rkapr.github.io | r.kapoor@tamu.edu | 979-xxx-xxxx | linkedin.com/in/rkapr | github.com/rkapr

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

EDUCATION:

Texas A&M University, College Station, TX
PhD in Electrical and Computer Engineering (GPA: 3.82) May 2017 – Dec 2021

Texas A&M University, College Station, TX
Master of Science in Electrical and Computer Engineering (GPA: 3.75) Aug 2015 – May 2017

Indian Institute of Technology, Patna, India
Bachelor of Technology in Electrical Engineering (GPA: 8.57/10) Aug 2010 – May 2014

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 maximum 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 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 (one step decision tree) as weak learner for detecting face images using vertical, horizontal and diagonal Haar features
- Vectorized code to speed-up repeated processing, observed effects of FP/FN penalty on empirical error
- ***Key achievement:*** Achieved **3.5x improvement** in training time at cost of 2.25% reduction in empirical accuracy by filtering robust features

Minimum description length (MDL) 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:*** Released MATLAB code based on Dougherty et al. paper was for open source use

Mixture of Poisson's for modeling number of daily deaths [Spring 2018]

- Derived, implemented and compared convergence of gradient descent and Newton Raphson optimization for maximum likelihood estimates
- Calculated *sympy* expressions for gradient and hessian in Python, then converted to *numpy* functions to avoid errors due to hard coding expressions

***Gaussian Mixture Modeling for Cancer Heterogeneit*y** [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

RESEARCH PROJECTS:

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

- Worked with professor from Soil and Crop Sciences Dept. to identify problem of interest
- Proposed and performed WGCNA based gene clustering using R scripts on computing cluster
- Identified UMAMI transporters in wheat, surveyed, collected, cleaned publicly available gene expression datasets for different development series, stress and tissue types
- Created heatmaps and homeolog expression plots for wheat and Arabidopsis to identify phylogenetically conserved patterns
- Developed shell based optimized pipeline to download large CRAM alignment files for relevant SRA projects using globus-cli, and organize runs for a given project
- Extracted reads mapping to UMAMI genes using SAMtools and BAMtools for gene model verification
- ***Key achievement:*** The analysis discovered two modules involved in activating storage molecules including starch, lipids and protein, discovered multiple transcription factors 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]

- Collaborated with professor from Dept. of Biochemistry and Biophysics to identify research problem
- Proposed Gaussian graphical modeling with fused penalty to uncover gene interactions
- Built end-to-end pipeline for cleaning, mapping, counting, DEseq analysis of RNAseq data
- Inferred causal gene interactions for sorghum circadian genes using unsupervised learning
- Researched literature to verify gene interactions of Arabidopsis, rice, maize orthologs
- ***Key achievement:*** Identified 21 potential regulators of sorgoleone biosynthesis genes, verified TF-gene interactions using 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/ AWARDS:

- Texas Engineering Experiment Station (TEES) Research Assistantship, Sept 2015 - present
- Texas International Student Scholarship, Fall 2015, Spring 2016