Lingfeng Huo

Saint Paul, MN • (612) 448-4122 • huo00030@umn.edu

EDUCATION

UNIVERSITY OF MINNESOTA, Minneapolis, MN
School of Statistics
Candidate for **Master of Science in Statistics** May 2023
First Term GPA (4.0): Advanced Regression Techniques(A), Theory of Statistics(A)
Courses in Progress: Design and Analysis of Experiments, Machine Learning: Analysis and Methods (in Python), Preparation for University Teaching

CHINA UNIVERSITY OF MINING AND TECHNOLOGY, Xuzhou, China
Bachelor of Science – Applied Mathematics June 2019
GPA:3.5 Relevant Courses: Time Series Analysis, Applied Stochastic Process, Visual C++ Programming

DATA SCIENCE PROJECTS

Statistical Modeling for Detecting Fraud claims in Auto insurance for Travelers | R

- Imputed the missing values for each predictor by multiple linear regression for a 17998×25 train set
- Transformed each predictor into normal variable by Box-Cox method; detected and deleted the cases with outliers by setting 4 as the z-score threshold
- Chose LASSO as the best variable selection method among stepwise method, MCP and SCAD under the criteria of VSD, F-measure and G-measure and selected 14 important predictors
- Trained models of LASSO, logistic regression with SOIL method, random forest and SVM, and compared them by F-score, Sensitivity, Specificity and Area under ROC curve with 10-fold Cross-Validation
- Used logistic regression to predict the response of test set with the best F-score among all team as 0.3733

Classification of 8×8 images by Machine Learning algorithms | Python

- Removed columns with excessive zeros and deleted cases containing outliers with z-score larger than 4 for a 1797×64 dataset transformed from 1797 images
- Used Fisher Discriminant to project the data onto two dimensions and employed bi-variate Gaussian generative model with parameters estimated by MLE to classify the data into 10 classes
- Used multiclass logistic regression to classify the data into 10 classes with the model parameters updated 100,000 times by steepest gradient method with the goal of minimizing the cross-entropy error function
- Used Naïve-Bayes classifier based on marginal Gaussian distributions to classify the data into 10 classes with model parameters estimated by MLE
- Chose logistic regression as the final model with the best average error rate as 0.14 and standard error as 0.015 in the comparison of the three classifiers through 10 pairs of train set and test set constructed by stratified random sampling

An Improved Optimization Method for BP Neural Network Training | MATLAB

- Created a hybrid optimization method called Conjugate Newton's Method (CNM) by combining Conjugate Direction Method and Newton's Method and proved its global convergence under exact line search
- Obtained an improved version of CNM called Quasi-Conjugate Direction Method (QCDM) by approximating the Hessian Matrix in CNM step with a positive definite matrix derived from Levenberg–Marquardt step
- Used 10-fold Cross-Validation to prove the network with ReLU as hidden layer activation function trained by QCDM has better R^2 than FR conjugate gradient method and steepest gradient method under different test functions

Design of Experiment for Discovering the Reasons of Choosing Take-out Food on Campus | R

- Designed a questionnaire that includes the potential predictors affecting takeout consumption and set monthly expenditure on takeout as the response; obtained a 324×5 data set
- Tested the normality of errors and homogeneity of error variances by normal Q-Q plot and residual plot; Applied Box-Cox transformation to the response to make the assumptions of errors hold
- Tested the significance of interaction effects and main effects by F-tests based on Type II Sum of Squares
- Found the combination of treatments that maximizes the takeout consumption by pairwise contrasts through Tukey HSD method; provided the student dining hall with advice of improvement based on the results

WORK EXPERIENCE

JD Technology (A Unit of JD.com), Guangzhou, China
Industrial Operation Intern June 2020 - September 2020

- Introduced the business model of JDT Digital Economy Industrial Park to enterprises, continuously addressed their needs and promoted 3 companies to enter the park
- Assisted the product manager to sell JDT's big data analytics model (C2B Reverse Customization System) to manufacturers in the park by explaining how the system can guide companies in product design by analyzing consumer's needs through JD.com's database and machine learning methods
- Analyzed the sales data of the JD stores in the parks of South China region during JD industry festival to provide guidance for the next industry festival's budget planning, advertising timeline and product selection

TECHNICAL SKILLS

Tools: Python, R, MATLAB, C++, SQL, SPSS, Excel, PPT **Techniques:** Statistical Modeling and Design of Experiment via R, Classification by Machine Learning Algorithm via Python