

Chenghao Ding

(217) 305-0997 ♦ frododch@gmail.com ♦ Web: chenghaoding90.github.io/ ♦ LinkedIn: linkedin.com/in/chenghaoding/

Education

University of Illinois at Urbana-Champaign

Ph.D. in Nuclear Engineering

Dec 2021

Master of Science in Applied Statistics

May 2020

Wuhan University

Master of Science in Power Engineering

Jun. 2015

Hubei University of Science and Technology

Bachelor of Science in Nuclear Engineering

Jun. 2013

Related Courses

Introduction to Data Science

Statistical Learning in Data Science

Advanced Data Science

Machine Learning

Applied Regression & Design

Statistics of Big Data & Clustering

Pattern Recognition

Statistical Learning

Experience

University of Illinois at Urbana-Champaign

Urbana, IL

Graduate Research Assistant

May. 2018 - Present

- Ph.D. Thesis: Global Heat Balance Model and Probability Distributions for Atmospheric Response
- Applied machine learning techniques (PCA, KNN) to anomaly detection
- Developed algorithms with KDE and MLE to find the probability distribution and optimized model parameters
- Applied SARIMAX model to study the correlated residuals and uncertainty quantification

Projects

Credit Card Fraud Detection (case study on Classification of Minority Class from Imbalanced Data)

Spring 2021

- Create a 50/50 sub sample ratio of "Fraud" and "Non-Fraud" transactions by using Near-Miss Algorithm
- Compare logistic regression with KNN, Decision Tree, and SVC with SMOTE technique, and achieve a f1 score of 0.93
- Neural Networks are created to compare the accuracy to the best classifier and obtain an accuracy score of 95.7%

CNN for Object Recognition in Images (case study on Fashion MNIST dataset)

Spring 2020

- 60,000 images and One-hot coding is used, and a CNN model is built by Keras and 97.99% test accuracy is achieved
- Tuning hyperparameters with skopt, the best learning rate, filter size of the convolutional layer and number of dense layer
- Test accuracy was improved to 98.36% and loss function converged quickly

COVID-19 County Level Data Analysis

Spring 2020

- County level data about demographics and health-related information are used to predict one week deaths
- Variables correlation with deaths are inspected and important features are selected
- K-means and agglomerative hierarchical clustering method are used to find Covid-19 growing pattern
- Random forest, Generalized additive model regression and XGBoost method are used, and weekly deaths are well predicted

Build a recommendation system on MovieLens 100K Dataset

Fall 2019

- Singular Value decomposition (SVD) method is applied, 5 fold cross-validation for training and testing are computed
- Implemented item-based collaborative filtering algorithm through Hadoop MapReduce chaining jobs
- Made movie recommendations based on top-k user rating predictions

Natural Language Processing on Disaster Tweets

Spring 2018

- Conducted data cleaning, exploratory analysis on 10,000 Tweets dataset
- Build a ML pipeline including bag of words, TF-IDF transformers, and XGBoost classifier, achieved prediction accuracy of 77%
- LSTM model combined with pre-trained GloVe 100D word net are built to retrain the dataset and improved accuracy to 81%

Skills

Computer

Python, SQL, TensorFlow, Keras, R, MATLAB, Java, Hadoop, Amazon Web Service

Data Science

NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, Plotly, XGBoost

Machine Learning

Predictive Modeling, Computer Vision, Recommender Systems