Yanwen LIN

+1-412-708-5446 | yanwenl@andrew.cmu.edu
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

Education Background

Dalian University of Technology Sep.2013 - Jun.2017
Bachelor of Engineering in Civil Engineering GPA: 3.84/4.00

Carnegie Mellon University Aug.2017 - Present
Master of Science in Civil and Environmental Engineering Current GPA: 3.91/4.00

Research Experience

Pittsburgh Fire Risk Modeling (on-going) Metro21 Inst., Carnegie Mellon University
Partnered with the City of Pittsburgh's Bureau of Fire Jun.2018 - present

- Explored historical fire dataset for Pittsburgh area, including time series analysis, text mining, etc.
- Designed consistency assessment framework for deployed fire risk model and evaluated its performance after incorporating weekly updated inspection and fire data

Bridge Placement Optimization for Rural Area Carnegie Mellon University
Partnered with Bridge to Prosperity Organization Jun.2018 – Jul.2018

- Cleaned and explored bridge assessment dataset, including correlation and geo-spatial analysis, etc.
- Designed a mathematical model to quantify the connectedness impact for the rural area and also did global sensitivity analysis on the model parameter using Sobol Indices via R

Audience Effect on Programmer Coding Style (on-going) STRUDEL Lab, Carnegie Mellon University
Supervised by Prof. Bogdan Vasilescu Feb.2018 - present

- Mined git repository data via python and pygit2 package and extracted features from raw data
- Queried specific GitHub data from the large GHtorrent database via SQL.

Independent Study: Urban Temperature Recognition Carnegie Mellon University
Supervised by Prof. Matteo Pozzi Mar.2018 – Jun.2018

- Analyzed temperature data correlation from both spatial and temporal perspective in order to build covariance function for Gaussian Process model.
- From temporal perspective, analyzed the autocorrelation of time history global temperature given an area.

Virtual Tracking Experimental System Based on Unity 3D Dalian University of Technology
Traffic Science and Technology Competition work Oct.2014 - Apr.2015

- Developed a virtual tracking experimental system programmed by C# in Unity3D
- Awarded the first prize in the 10th Traffic Science and Technology Competition in DUT

Course Project

Course: 15-619 Cloud Computing Carnegie Mellon University

Big Data Analysis on Wikipedia Dataset with MapReduce Sep.2018

- Developed a pre-processing workflow for Wikimedia Pageview dataset
- Applied workflow to a 320GB dataset with AWS Elastic MapReduce Hadoop service under limited budget.
- Analyzed processed data with Jupyter notebook on remote server using Pandas

Course: 15-688 Practical Data Science Carnegie Mellon University

Cannes Film Festival Prediction based on Twitter Analysis Apr.2018 – May.2018

- Developed a pipeline for data collecting, pre-processing and modeling to predict 2018 Cannes winner list.
- For machine learning, to tackle dataset imbalance, we constructed several models such as ensemble, SVM and logistic regression to do cross-validation based on AUC score for model selection.

Spam Email Recognition and Collaborative Filtering Apr.2018

- Constructed an RBF SVM model and did cross validation to filter spam email from around 10K emails
- Built a recommendation system using collaborative filtering from a 100k ratings of MovieLens dataset

Tutorial on Latent Dirichlet Allocation Mar.2018

- Described thoroughly the probabilistic graphical model of LDA .
- Elaborated the implementation of LDA model, specifically via Standard and Collapsed Gibbs sampling.

Text Classification and Natural Language Processing Mar.2018

- Tokenized raw Twitter text using NLTK package, extracted features from it and did text classification.
- Implemented a class to process raw Federalist Papers, including both TFIDF and N-gram language models.

Data Collection via Web Scraping and Parsing Implementation Jan.2018

- Used python (BeautifulSoup package) to scrap HTML of Yelp source page to get useful restaurants information.
- Implement a parser class to parse practical XML file

Course: 10-701 Introduction to Machine Learning (PhD level) Carnegie Mellon University

Hand-write Digit Recognition Oct.2017 – Nov.2017

- Implemented a CNN for digit recognition including forward and backward using MATLAB.
- Clustered unlabeled digit image via writing EM algorithm after PCA dimension reduction

Applied Machine Learning for Gene Expression Profile Classification Nov.2017 - Dec.2017

- Performed dimensionality reduction on real cell gene expression profile data (~20 thousand dimension) using PCA/kernel PCA, LDA, etc.
- Predicted the types of certain cells based on ensemble of several basic algorithms including logistic regression, SVM, neural network, Gaussian Process, etc.

Skills

- **Technical skills:** Data science/mining, machine learning, database system, Linux(familiar), team working
- **Programming language**: Python (NumPy, Pandas, Scikit-learn, etc.), Java (Hadoop, maven), R, C, MySQL
- **Application software**: Jupyter Notebook, MATLAB, Latex, ABAQUS, AutoCAD