Some of the MS in Data science projects

- Improve Image Classification Using Data Augmentation and Neural Networks

  Jan 2019 – Aug 2019

  Project description

  Model architecture, hyper-parameter tuning, and data augmentation are essential to reduce model overfitting and help build a more reliable convolutional neural network model. The model performance with CIFAR-10 image dataset (60,000 32x32 color images in 10 different classes) was evaluated with accuracy and loss, characteristics derived from the confusion matrices, and visualizations with non-linear mapping algorithm t-SNE. The complete VGG16 model achieves train accuracy at 96% and test accuracy at 92% with values of loss function less than 0.5.

- NLP Sentiment Analysis of IMDb Movie Reviews

  May 2019 – Aug 2019

  Project descriptions

  - Use BeautifulSoup to compile 956 individual user movie review permalinks from IMDb, preprocess and normalize corpus, vectorize with Bag of Words (BoW) or Term-Frequency-Inverse-Document Frequency (TF-IDF)

  - Scrape Children's books from Gutenberg Project website using BeautifulSoup, preprocess to clean licensing and preface of the books, Topic model pipeline: Removal of accented characters, contraction expansion, making text in lowercase, remove extra newlines and spaces, text lemmatization(SpaCy) , special character removal, stop words removal (SpaCy) to normalize corpus, vectorize with Bag of Words (BoW) or Term-Frequency-Inverse-Document Frequency (TF-IDF), compare K-Means and hierarchical clustering models with Silhouette coefficient, uncover hidden structures with topic modeling (NMF: Nonnegative Matrix Factorization, LDA: Latent Dirichlet Allocation) and visualization with pyLDAvis.

- Use R and Python for Case Studies in Quantifying the World

  Jan 2019 – May 2019

  Project description

- Apply R for case study: (1) Explore MAC addresses to predict real-time location with weighted and un-weighted k-NN methods; (2) Model the change of run time for individuals in Cherry Blossom 10-mile run with LOESS fit curves; (3) Compare Type I and Type II errors in identifying spam emails.
- Apply Python for case study: (4) Study volatility and return for US technology portfolio using momentum stock trading strategy with signal frontier analysis; (5) Compare different types of missing data (MCAR, MAR or MNAR) imputed with mean and MCMC methods; (6) Analyze Higgs data with neural networks tuned in different architectures and hyperparameters.
- Essay: (7) Artificial Intelligence is transforming healthcare and life science industries.

- Case Studies for Machine Learning

  Jan 2019 – Mar 2019

  Project description

  Use Python Spyder to conduct these case studies: (1) Set up grid search function for large number of classifiers and hyperparameter settings; (2) Investigate specific claims in a large dataset (500,000 instances, 29 attributes), create best models with GridSearchCV for claim prediction, and find the predominately influencing attributes; (3) Recommendation system and decision making with matrices.

- Online News Popularity Prediction

  Sep 2018 – Dec 2018

  Project description

  Predict the popularity of a news article using the online news popularity dataset from UCI Machine Learning Repository (39,797 instances, 61 attributes). Recursive feature elimination (RFE) is used for feature selection during data preprocessing, different regression and classification models (e.g. linear regression, logistic regression, SVM, random forest, decision tree learning, Gaussian NB, Gradient Boosting) are compared with model accuracy, parameters are optimized with GridSearchCV, and pipelines are constructed to combine PCA and KNN, RF or Gaussian NB classification method. Dimensionality reduction techniques (i.e. PCA, t-SNE) are used in different clustering models (e.g. K-Means, spectral, hierarchical) which were further evaluated with Silhouette Coefficient analysis.