Stat 766 Applied Data Mining/Machine Learning and Predictive Analytics

Syllabus

Department of Statistics Kansas State University

Fall of Odd Years



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Course information

Course: STAT 766 Applied Data MiningMachine Learning

and Predictive Analytics

Lectures: T U 8:05 - 9:20 a.m., Dickens Hall 106 Instructor: Haiyan Wang, Department of Statistics

Office: 102A Dickens (tel: 532-0524)

Email: hwang@ksu.edu

Prerequisite: STAT 705 or STAT 713 or STAT 717, and prior com-

puter programming proficiency on C, C++, Fortran,

R, or Python (e.g., CIS 209, STAT 726)

Office Hours: MWF4:30 - 5:30 pm or by appointment.

Course link: Sign in from K-State Online

Course description

This course will learn the complete process of building analytical tools suitable for learning from text, image, and numerical data, including

- Data scraping and wrangling: automatic online data collection, feature extraction. Automatic retrieval of various format online data, including JSON, REST, and Streaming API, http(s), html, xml, and databases.
- Methods: state of the art supervised and unsupervised statistical machine learning methods, including regularized linear and logistic regression, classification trees, nearest neighbor methods, support vector machines, nave Bayes, random forests, boosting/bagging/AdaBoost, clustering, statistical text processing/mining, topic modeling models such as latent Dirichlet allocation, convolutional neural network analysis and various network structures.
- Case studies and hands on applications: image category classification, predicting stock market returns, detecting fraudulent transactions, analyzing sentiment of product reviews, recommender system, mining social media and news data for intelligent insight.
- Evaluation, model deployment, and report writing.



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Chapter 1: Python 3 basics

- Python 3 data structures
 - List, dictionary, tuples, Pandas DataFrame, NumPy array, tensors
 - Work with string and files in Python 3
 - String in Python 3
 - Work with files
- SQLite Relational Databases and sqlite with python

Chapter 2: Part I: Web scraping to get data

- Intro to web crawling
 - BeautifulSoup and urllib.request
- Web scraping with selenium webdriver
 - Intro to selenium webdriver
 - Find elements for selenium webdriver
 - Navigation and frames
 - Selenium wait commands
- Example: Scrape COVID references with selenium webdriver
 - Find elements by class name
 - Find elements by xpath
- Example: Scraping Reuters news data
 - Getting title and url of news articles
 - Writing to database while scraping
- Getting patient characteristic data from GSE11417



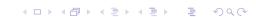
Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Chapter 2: Part 2 Getting data with REST API

- REST API Introduction
 - JSON and python command for JSON
 - Status code
 - How to use API
- Use Python to make REST API Requests
 - Example: Request business infor with Yelp Fusion API
 - Example: Request data with Twitter API
 - Getting real time streaming tweets

Chapter 3: Text preprocessing and word embeddings

- Pre-processing
 - Tokenization
 - Remove stop words
 - Remove punctuation characters
- Feature extraction
 - Document-term matrix (DTM)
 - Term Frequency Inverse Document Frequency (TFIDF)
- Regularized Logistic Regression
 - Logistic Regression for Binary classes
 - Multinomial Logistic Regression
- Word embedding
 - Word2Vec
 - Glove
 - FastText



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Chapter 4: Model evaluation

- Performance Measures for Categorical Targets
 - Basics
 - Measures for binary classification
 - Measures for multinomial targets
- Performance Measures for Continuous Targets
 - Basic Measures of Error
 - Domain Independent Measures of Error
 - Bias and variance trade-off
- Designing Evaluation Experiments
 - Hold-out Sampling
 - Cross Validation
 - Bootstrapping
 - Case study: IMDB movie review classification
- Evaluating Models after Deployment
 - Monitoring Changes in Performance Measures
 - Monitoring Model Output Distributions
 - Monitoring Descriptive Feature Distribution Changes

Chapter 5: Other traditional supervised and unsupervised models

- Supervised learning models
 - Nearest neighbor methods
 - Nave Bayes and PAM
 - Tree and random forests
 - Boosting/bagging/AdaBoost
 - XGBoost
- Unsupervised methods
 - Dimension reduction with PCA, SVD, and TSNE
 - Clustering methods for numerical data and topic modeling models (such as latent Dirichlet allocation) for text data



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Chapter 6: Gradient descent basics

- Gradient descent
 - One dimension
 - Multiple dimensions
 - Application to SVM objective
- Stochastic Gradient Descent
 - Basic SGD and mini-batch SGD
 - RMSprop, Adagrad, Adadelta, Adam
 - Application to ridge regression
- Coordinate Gradient Descent
 - Basics and convergence
 - Application in Lasso



Chapter 7: Margin maximization and Support vector machine

- Maximizing the margin for binary classification
 - Rewrite the objective function
- Support vector machine
 - Solving SVM objective with quadratic programming
 - Using gradient descent to solve SVM objective
 - Insight
- Extensions to SVM with nonlinear kernels



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Chapter 8: Neural network, RNN, & Convolutional NN

- Neural networks
 - Basics of feed-forward Neural Networks
 - Single layer
 - Multiple layers
 - Choices of activation function
 - Error back-propagation
 - Loss functions and activation functions
 - Two-class classification and log likelihood
 - Multi-class classification and log likelihood
 - Alternative considerations for optimizing parameters
 - Mini-batch gradient descent
 - Adaptive step-size
 - Regularization
- Recurrent neural network (RNN)
 - LSTM
 - Application in image analysis
 - Application in natural language processing
- Convolutional Neural Networks
 - Nature of CNNs
 - Filters
 - State of the art networks on computer vision: AlexNet, GoogleNet, VGG, ResNet
 - Implementation and application



Chapter 9: State of the art models for NLP

- Deep average network (DAN)
 - Nonlinear activation functions
 - Specs and performance of DAN
- Transformer models
 - Attention
 - Components of Transformer model
 - OpenAl Transformer
- BERT
 - BERT motivation and architecture
 - BERT model strategy
 - Summary of BERT and references
- Universal Sentence Encoder
 - Motivation and USE architecture
 - Parameter estimation in USE
 - Performance of USE on transfer tasks
 - Application of USE



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

References

- Sebastian Raschka and Vahid Mirjalili Python (2019) Machine Learning Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing.
- Anubhav Singh, Sayak Paul (2020) Hands-On Python Deep Learning for the Web: Integrating neural network architectures to build smart web apps with Flask, Django, and TensorFlow. Packt Publishing.
- Nikhil Ketkar, Jojo Moolayil (2021) Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch. Apress.
- François Chollet (2018) Deep Learning with Python. Manning.
- Python Machine Learning for Beginners: Learning from Scratch Numpy, Pandas, Matplotlib, Seaborn, SKlearn and TensorFlow 2.0 for Machine Learning & Deep Learning- With Exercises and Hands-on Projects. (2020) Al Publishing LLC.
- Taweh Beysolow II (2018) Applied Natural Language Processing with Python.
 Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing. Apress.
- Nikhil Ketkar (2017) Deep Learning with Python. A Hands-on Introduction. Apress.



References

- Convolutional Neural Networks in Python: Master Data Science and Machine Learning with Modern Deep Learning in Python, Theano, and TensorFlow (Machine Learning in Python). (2016) LazyProgrammer
- Navin Kumar Manaswi (2018) Deep Learning with Applications Using Python:
 Chatbots and Face, Object, and Speech Recognition with Tensorflow and Keras.
- Palash Goyal, Sumit Pandey, Karan Jain (2018) Deep Learning for Natural Language Processing. Creating Neural Networks with Python
- Adrian Rosebrock (2017) Deep Learning for Computer Vision with Python.
 2-Practitioner Bundle. PylmageSearch.
- Adrian Rosebrock (2017) Deep Learning for Computer Vision with Python.
 3-ImageNetBundle. PyImageSearch.



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

References

- Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie (2008) The Elements of Statistical Learning. Second edition. Springer. https:
 - //web.stanford.edu/~hastie/Papers/ESLII.pdf
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) An Introduction to Statistical Learning: with Applications in R. Springer New York.
- Luis Torgo (2017) Data Mining with R Learning with Case Studies, second edition, Chapman & Hall/CRC.
- Richard Heimann, Nathan Danneman (2014) Social
 Median Mining with R: Deploy cutting-edge sentiment
 analysis techniques to real-world social media data using
 R. PACKT publishing.

Grading policy

There are assignments and final project. Assignments account for 60% and the final project accounts for 40% of the course grade. Grading scale

- [88, 100] A Excellent
- [77, 88) B Passing
- [60, 77) C Needs improvement
- [0,60) D



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Assignments

- Each assignment consists of two parts:
 - A. Write and submit your answer to the questions in the assignment.
 - B. On assignment due date/time, each student who submitted their answer will be randomly assigned to review two other students' answers. Write and submit your peer reviews within three days after the assignment due date.

Part B is meant for the following purpose:

- Go over the course content while your memory is fresh
- Learn from your peers and improve your skill set by analytically comparing your answers with those of peers
- All submissions should be online electronically
- You must submit pdfs + source files for part A, and word or pdf file for part B.

Assignments -conti.

Grading criteria

- (1) Did you answer the scientific question? (30%)
- (2) Did you use appropriate approach(s)/method(s)? (40%)
- (3) Was your write-up simple, clear, and precise? (20%)
- (4) Was your code reproducible? (10%)

Instruction for peer review:

- The peer review should include
 - a brief summary of what approach(s) the student took to answer each assignment question;
 - how well that appoach(s) works;
 - any major and minor problems with the approach(s);
 - any suggestion to improve.
 - a score for each of above four grading criteria
- Peer review is due three days after it is assigned.

Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Final Project

- The final project will be longer than the weekly assignments in terms of required depth and thoroughness in analysis and writeup.
- You will have an opportunity to submit your analysis, get feedback from peers and the instructor, and re-analyze the data on this project.
- I will give you an option of a final project, but you may also propose a different project to the instructor and as long as it is approved you can proceed with that project. The project should involve data/code that you write/obtain, process, analyze, and synthesize yourself.
- The project proposal is due on Nov 9, 2022.

Project proposal for the final project

The proposal should include the following:

- Statement of the problem
- Significance of the problem. Why is it important?
- What is the target conclusions/outcome of the project?
- Literature review of what others have done and what are the state of the art solutions to the problem?
- Describe how you plan to approach the problem, and give the timeline of your planned step-by-step activities.
- Preliminary analysis result to show that your planned approach will work.



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana

Structure of Class Time

- Class will consist of both lectures on methodology and practice, as well as hands on activities.
- It is critically important that you go over the lecture slides and try the code in the slides before each class.
- The examples in slides are best to be tried by yourself in advance of each lecture so that you have time to look it over and come up with questions.
- Warning: If you don't go over the slides and try examples ahead of time, you will have a hard time to keep up with the content.

Data mining/machine learning is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.

The following are related to each other (some are interchangable and some are sub-field of the bigger category):

- Data science
- Data mining
- Machine learning
- Artificial intelligence
- Natural language processing
- Signal processing
- Business analytics
- Text mining



Haiyan Wang: Stat 766 App. Data Min./Mach. Learn.& Pred. Ana