DS 7333 Quantifying the World

Course Overview

Quantifying the World (QTW) is designed to cover multiple machine learning and statistical methods primarily in two-week chunks. There will be approximately 90 minutes of background material and 90 minutes of live instruction per week. There will be a case study utilizing the methods learned due every two weeks. The topics included in this course are supervised and unsupervised methods of learning, linear and nonlinear classifiers, and model selection and hyperparameter tuning. The overall goal of the course is to put all of the pieces of data science learning in the MSDS program together and build best practices for developing statistical models.

Required Text

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2017. *The Elements of Statistical Learning*. http://web.stanford.edu/~hastie/ElemStatLearn.

Learning Objectives

Students will learn to:

- Create and train both supervised and unsupervised models
- Create and train linear and nonlinear models
- Learn the scaling abilities of multiple modeling methods
- Learn to properly tune the hyperparameters of a model
- Become familiar with the state-of-the art algorithms in practice today
- Apply statistical models to a variety of datasets

Assignment/Assessment Weight on Final Grade:
Case Studies 50%
Participation 30%
Quizzes 20%

Case Studies (50%): Every other week in class will be spent working through a case study using Python. These case studies will be written up and turned in for a grade. Case study write-ups are to be “technical report ready,” which basically means that it is suitable for public viewing.

Each case study will be graded out of 100 points. The final case study will count as a double grade and serves as a final project. See the table below for case study due dates. Each case study has a corresponding video element in the asynchronous material.

Participation (30%): Each week after viewing the asynchronous materials, students are to upload a one- to two-page Word or PowerPoint file with questions or comments about the subject matter of the week. Submit to the week 1–14 Presession Question Submission.

Quizzes (20%): Each week has a quiz on the asynchronous material. Quizzes must be submitted 24 hours after the live session.

Weekly Schedule

Week	Topic(s)	Case Study
1: Introduction to Python	Course Introduction Python Linear Regression	
2: Imputation	Missing Data Patterns Imputing Data Cross-Validation	
3: Logistic Regression	Loss Functions Regularization Logistic Regression	Case study 1 (covers weeks 1–2). Due by midnight CT on the day of the live session.
4: Regularization	Feature Selection Overfitting Multiclass Classifiers	
5: Clustering	Unsupervised Learning Clustering Distance Metrics	Case study 2 (covers weeks 3–4). Due by midnight CT on the day of the live session.
6: Naive Bayes	Naive Bayes Bag of Words Models	
7: Decision Trees	Entropy, Gini, Information Gain Partition Trees Random Forests	Case study 3 (covers weeks 5–6). Due by midnight CT on the day of the live session.
8: Boosting	XGBoost	
9: Vector Machines	Support Vector Machines Kernel Trick Hinge Loss	Case study 4 (covers weeks 7–8). Due by midnight CT on the day of the live session.
10: Out of Core Methods	Stochastic Gradient Descent Vowpal Wabbit Hashing	

11: Introduction to Neural Networks	Basic Neural Networks Dense Layers	Case study 5 (covers weeks 9–10). Due by midnight CT on the day of the live session.
12: Convolutional Neural Networks	Convolutional Neural Networks Image Recognition	
13: Recurrent Neural Networks	Recurrent Neural Networks NLP	Case study 6 (covers weeks 11–12). Due by midnight CT on the day of the live session.
14: Ensembling	Ensembling Random Search Grid Search	
15: Model Deployment	Docker, APIs, and Deploying Models	Final case study (covers weeks 13 and 14; note that week 15 material is not required for Final Case Study Submission)