NYU Big Data Course Syllabus 2025H2

Overview

This course will focus on several aspects of model building and evaluation. Through hands-on practice, students will learn how to process large datasets, how to construct useful models, and how to evaluate the quality of those models.

The course will discuss many practical problems that are rarely seen in other courses. The course will also review some theoretical aspects of the problem space, but no particular mathematical or statistical prerequisites are needed, we will provide the necessary background.

It is helpful to know how to program in at least one language, but we will provide a good deal of sample code and actual programming will be relatively light.

Schedule

Week 1: (Ken)

- Introduction to AI, Machine Learning, & Data Science
  - General Introduction
  - Basic concepts
  - Machine Learning Models
  - NLP
  - Generative AI
  - Python/R/Other tools
- Homework:
  - Knowledge review and Python or R coding exercise

Week 2: (Tyler)

- Introduction to Apache Spark
  - Setting up
- Introduction to entropy
  - Shannon's understanding of entropy
  - Modeling signal vs. Noise
  - KL-divergence as a concept
  - Pathologies as inaccurate entropy estimations
- Homework: First project using Apache Spark
  - Fit simple model on example data set. Compute entropies.

Week 3: (Ken)

- Supervised Learning Models (I)
- Modeling skills / fine tuning
    - Overfit vs underfit
    - Regularization
    - Cross validation
    - Optimization
- Homework:
    - Design & build supervised learning model(s) in Python or R

Week 4: (Tyler)

- Classification Models
    - Multinomial Logistic classification model using cleaned mortgage data
    - Explain the data set, explain the model.
    - Simple theoretical results in this space, connections to entropy.
- Homework: Classification model assigned
    - Build Classifier in Spark using multi-logit regression and random forest.

Week 5: (Ken)

- Supervised Learning Models (II)
- Model Evaluation & Comparison
    - Confusion Matrix
    - Precision / Recall Curve
    - ROC / AUC
    - Other evaluation methods
- Homework
    - Develop and Evaluate classifiers in Python or R

Week 6: (Tyler)

- Neural Networks
    - Basic structure
    - Examine results from “pace car” model
    - Explore information criteria, AIC, BIC, TIC
    - White’s criteria
- Homework
    - Fit a neural net model.

Week 7: (One of us)

Midterm test

Week 8: (Ken)

- Unsupervised Learning Models (Clustering & Pattern Recognition)
    - K-Means
    - Hierarchical
    - Association Rules
    - Others

- Homework
  - Application of clustering techniques in Python or R

Week 9: (Tyler)

- Image Classification
  - Basic structure, basic principles.
  - Convolutional Neural Networks
- Homework
  - Build your first CNN model for the MNIST dataset.

Week 10: (Ken)

- Natural Language Processing & Deep Learning
  - Parsing & tokenization
  - Conversion of text into digital format
  - Sentiment Analysis
  - Classic DL models
- Homework
  - Design and build a DL based NLP solution in Python or R

Week 11: (Tyler)

- Model Pathologies and Overfitting
  - Examine the ways in which models go wrong.
  - Why do some of our algorithms not work, what breaks them.
  - The right way to do feature selection.
  - The right way to do model selection.
- Homework
  - Try to select the best of our previous models to run on unseen data.

Week 12: (Ken)

- Generative AI & LLM techniques
  - LLM model introduction
  - Prompt engineering
  - Chunking & embedding
  - RAG / Hybrid RAG
  - Financial LLM use cases - trading strategies / chatbot / summarization etc.
- Homework
  - Design & build a simple LLM based solution

Week 13: (Tyler)

- Word2Vec deep(er) dive.
  - We will examine word2vec, one of the oldest and simplest embedding models.
  - Discuss embedding models, what they are, how they work.
  - Examine common problems in current embedding models and their associated large language models.
- Homework
  - Build some simple examples with Word2Vec.

Week 14: (either)

- Review homework, prepare for final exam/project.
- Review current state of the art and research in this field.

The NYU Tandon School values an inclusive and equitable environment for all our students. I hope to foster a sense of community in this class and consider it a place where individuals of all backgrounds, beliefs, ethnicities, national origins, gender identities, sexual orientations, religious and political affiliations, and abilities will be treated with respect. It is my intent that all students' learning needs be addressed both in and out of class and that the diversity that students bring to this class be viewed as a resource, strength, and benefit. If this standard is not being upheld, please feel free to speak with me.