NYU Big Data Course Syllabus 2018/01

Week 1: (Ken)
- Introduction to data science
  - Data science concepts
  - Statistics concepts
  - Classification & clustering
  - Introduction of R
- Homework:
  - Statistical analysis and data visualization in R / Python

Week 2: (Tyler)
- Introduction to entropy
  - Shannon’s understanding of entropy
  - Modeling signal vs. Noise
  - Entropy allocation between parameters and residuals
  - 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)
- Regression Models
  - Linear Regression
  - Data transformation for Regression
  - Non-linear Regression / Logistic Regression
- Various Concepts in Machine Learning Modeling
  - Overfit vs underfit
  - Regularization
  - Cross validation
  - Optimization
- Homework:
  - Design and build a regression model in R / Python

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.

Week 5: (Ken)
- Classic Classifiers
  - Decision Tree
- Naive Bayes
- SVM
- Model Comparison & Evaluation
- Use Cases
  - Homework:
    - Design and build classifiers in R / Python

Week 6: (Tyler)

- Model Pathologies and Overfitting
  - Examine results of previous round of classifiers.
  - Explore some common model mishaps
  - Examine results from “pace car” model
  - Explore information criteria, AIC, BIC, TIC
  - White’s criteria
  - Homework
    - No homework, prep for midterm.

Week 7: (One of us)

Midterm test

Week 8: (Ken)

- Clustering & Pattern Recognition
  - K-Means
  - Hierarchical
  - Association Rules
  - MCL
  - Homework
    - Application of clustering techniques in R / Python

Week 9: (Tyler)

- Introduction to cloud computing.
  - Cloud computing with google cloud.
  - Computational costs, accuracy, performance.
  - Compare known models, compare models from literature.
  - Homework
    - Classification model on the cloud

Week 10: (Ken)

- Natural Language Processing
  - Parsing & tokenization
  - Conversion of text into digital format
  - Classification & clustering
  - Sentiment Analysis
  - Homework
• Design and build a NLP model in R / Python

Week 11: (Tyler)

  o First neural nets
    • Examine basic theory and practice of neural nets.
    • Kolmogorov-Arnold representation theorem, and various analytic approximations.
    • Common pitfalls.
  o Homework
    • Build Neural Net using spark, compare to previous rounds of semi-parametric models.

Week 12: (Ken)

  o ML & Big Data Applications
    • Retail Banking
    • Marketing / Call Center
    • Mortgage Risk
    • Lead Generation
  o Homework
    • Review what has been taught in the whole semester

Week 13: (Tyler)

  o Deep learning models
    • Examine image recognition as a problem, look at neural nets built for this problem.
  o Homework
    • Build a CNN to perform image recognition.

Week 14: (either)

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