Machine Learning Intro

Spencer Lyon

Machine Learning

Acknowledgments: Materials (loosely) based on

- Course co-taught with Georgios Anagastoplos
- "Business Data Science" book and associated course slides by Matt Taddy

Goal

- Goal of this course is to give you the tools to turn messy real-world data into actionable insight directly relevant to business (or policy) decision making
- We will blend knowledge of programming, data know-how, and machine learning
- We'll call this business data science (borrowed term from Taddy)

Approach

- We will develop and practice a consistent approach for doing business data science
- Approach will combine data and models
- Inputs include data, domain knowledge, computer code, model specification, training algorithm
- Outputs include parameter estimates, metrics, graphs/charts, and recommended actions or responses

Data

- Population: A domain from which one can sample data
- Data generating process: the physical process generating the population
- Sample: an observation or data point drawn from the population
  - Indexed by i
  - Often represented as input, output pairs: $(\mathbf{x}_i, y_i)$
  - Input space $\mathbb{X}$ called feature space
  - Output space $\mathbb{Y}$ called target space (also label, or output)

Splitting Data

- Training data: Used in inverse problem (fitting or finding coefficients)
- Validation data: Used for model validation (to be explained later)
- Test data: Used to assess model performance
- Hold out data: Data not used for training (validation + test data)

Models

- Models tie data to outcomes using parameters
- We'll represent parameters by a vector $\theta \in \Theta$
- Given data (X, y) a model $f : \mathbb{X} \times \Theta \Rightarrow \mathbb{Y}$
- The model $f(x; \theta)$

Spectrum of Modeling Approaches

- There are many fields that study statistical models
- These fields can be loosely placed on a spectrum:
  Econometrics $\rightarrow$ Statistics $\rightarrow$ Data Mining/Data Science $\rightarrow$ Machine Learning $\rightarrow$ Deep Learning + AI
- This spectrum also aligns with a spectrum of goals/intents
  Measurement $\rightarrow$ Causality $\rightarrow$ Prediction $\rightarrow$ Accuracy
- All models should be constructed based on an understanding of measurement process, causal structure, and predictive capacity
- Different fields (and their algorithms) prioritize different parts of the spectrum

Algorithms: How Your Machine "Learns"

- The "learning" part of machine learning is the process by which parameters are fit so that the model can perform its task
  - *Note:* This is solving the inverse problem
- Many classical algorithms come directly from statistics or mathematics and are appropriate for a variety of tasks (OLS, SVD, PCA)
- As data gets large (in number of dimensions and/or observations), classical methods become intractable
- Many advances in algorithms over the past 15 years have extended the boundaries of tractability and pushed ML into new domains

Workflow: Progressive Complexity

- Start as simple as possible: e.g. sample moments
- Evaluate key metrics/targets using current stage model
  - Learn what works in model for data + domain + target
- Add features/complexity/model *power* to form next model
- Evaluate relative to benchmark of previous models
  - If not improving, re-evaluate structure of more complex model
- Know when to stop!

Example Workflow

1. Exploratory data analysis (charts)
2. Copy models (tomorrow looks like today, or tomorrow looks like that day last week)
3. Simple moment models (Moving average of past 7 days, hour by hour)
4. Linear Regression
5. Other linear ML
6. Time series models
7. Weighted time models
8. Non linear ML
9. Not so deep learning
10. Deep learning

PyData

We will continue to make use of PyData libraries

- Numpy
- Scipy
- Matplotlib
- Pandas

Machine Learning

We will also learn some new tools, specialized for machine learning

- Scikit-Learn
- Tensorflow
- PyTorch