Machine Learning Introduction

Spencer Lyon

Machine Learning Introduction

Course Objective

- 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
- By the end of the course you should have a solid foundation in both the theory/concepts behind modern machine learning (including deep learning) as well as the practical skills to put the theory to work

What is Machine Learning?

- Machine Learning is...
  - The study and application of algorithms that improve by repetition and experience
  - A blend concepts from computer science and statistics
  - Typically used to solve the inverse problem for a particular statistical model (determining optimal parameters of a specific model, based on observed data)
- Machine learning isn't...
  - Magic: does require effort from user
  - Harmless: bias, censoring

Relation to Statistics (and Econometrics)

- 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
- ML typically prioritizes prediction and accuracy

Families of Machine Learning

- Supervised Learning: learn to map features to targets (labels)
  - Regression: targets in $\mathbb{R}^N$
  - Classification: targets in discrete space
- Unsupervised learning: discover structure without labels
  - Clustering
  - Dimensionality reduction
  - Compression
- Reinforcement Learning: learn by doing
  - Learn optimal behavior by interacting with environment
  - Observe $\Rightarrow$ act $\Rightarrow$ rewarded $\Rightarrow$ observe...

Key Ingredients

1. Data
2. Model
3. Algorithms/Estimation Procedure
4. Communication: key ”soft skill” for your work to have impact

1. 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)

2. Model

- 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 $f(x; \theta)$ generates predictions (for supervised learning) or performs another desired task

3. 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

4. Communication

- Last (but certainly not least) we have communication
- The algorithms you will be developing have great power...
- "but with great power comes [the] great responsibility" to explain the model and its implications to others
- Being an effective communicator is the only way your models can have an impact on key business or policy outcomes

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

Tools: PyData

We will continue to make use of PyData libraries

- Numpy
- Scipy
- Matplotlib
- Pandas

Tools: Machine Learning

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

- Scikit-Learn
- Tensorflow
- PyTorch