# An Introduction to Machine Learning for Social Scientists

Tyler Ransom

University of Oklahoma, Dept. of Economics

November 10, 2017

# Outline

# What is machine learning? What is AI?

- **Machine learning (ML):** Allowing computers to learn for themselves without explicitly being programmed
  - **USPS:** Computer to read handwriting on envelopes
  - **Google:** AlphaGo, computer that defeated world champion Go player
  - **Apple/Amazon/Microsoft:** Siri, Alexa, Cortana voice assistants
- **Artificial intelligence (AI):** Constructing machines (robots, computers) to think and act like human beings
- ML is a subset of AI

# ML in the social sciences

- A branch of statistics devoted to accurate prediction
- Maximize both in- and out-of-sample prediction
- Systematically combine estimation and model selection
- Computational techniques for stats on very large data sets
- Becoming more popular in “big data” era
- These slides based in part on Varian (2014)

## Motivating example

- Suppose you want to predict mortgage loan default (0/1 outcome)
- You have a large number (over 5,000) of relevant *variables*
- What would you do?
- There are better methods of prediction than logit:
  - Help you determine which of the 5,000 variables are most important
  - Automatically detect interactions among variables
  - Do a better job of predicting out-of-sample than logit

# Overfitting

- **Overfitting:** estimating a model that performs well in-sample but poorly out-of-sample
- **Example:** Suppose you have cross-sectional data for a continuous outcome across *n* individuals
- One way to predict earnings is to use OLS and estimate *n* dummy variable coefficients (no constant)
- $R^2 = 1$, indicating perfect in-sample fit
- But if I gave you a separate sample of this data with *m* different individuals, how would you predict the outcome? Which dummy coefficients would you assign to the new individuals?

# Solution to overfitting

1. Penalizing parameter complexity (Adjusted $R^2$, AIC, BIC)
2. Testing a variety of models out-of-sample
3. Using cross-validation to find the best level of penalty

# How cross-validation works

Typical steps used to cross-validate and test predictions:

1. Randomly divide up your data into three parts: training set (60%), cross-validation set (20%), and test set (20%)
2. Estimate your model parameters in the training set
3. Compute the prediction error in both the cross-validation and test sets
4. Repeat this for various levels of penalty
5. Pick the penalty level that minimizes error in the cross-validation set
   - Test set should only be used for out-of-sample prediction; some people lump test/CV together

# Outline

# Commonly used machine learning algorithms

- Continuous dependent variable:
    - Ordinary least squares
    - Regression trees / random forests
    - Penalized regression (LASSO, Ridge, Elastic net)
    - Nearest neighbor
    - Support vector machine (SVM)
    - Neural network
    - Naive Bayes
- Categorical dependent variable:
    - Logistic regression
    - All others above

# Ensemble prediction

- Often times, you will obtain better prediction by averaging across models (e.g. forests vs. trees; Bajari et al. (2015))
- e.g. obtain predictions from Penalized logistic regression, classification tree, and support vector machine
- Create a meta-prediction by regressing (in the cross-validation set) the outcome on the predictions from each model
- The meta-prediction will usually perform better in the test set than any single prediction
- But it's harder to back out the decision rule from meta-predictions

# Software to estimate ML models

- R and Python are the home of machine learning development
- Growing community in Julia
- Matlab has a ML toolbox, but lacks customizability
- Limited availability in Stata

# Unsupervised learning

- Up to now, we've only discussed *supervised* learning
- *Unsupervised* learning $\Rightarrow$ no dependent variable
- Used primarily to reduce large datasets
- e.g. detect partitions in data (*k*-means clustering, EM algorithm)
- Reduce dimensionality of data (PCA)

# Outline

1. Intro
2. Examples
3. Conclusion

# Limitations of machine learning

- Machine learning is all about *prediction* (i.e. correlation)
- But social science is primarily motivated by *causality* (i.e. prediction in a counterfactual environment)
- Attempts currently being made to re-frame machine learning in terms of causal inference (Varian, 2014; Athey and Imbens, 2015; Bajari et al., 2015)
- Or to detect groups of unobserved heterogeneity using unsupervised ML (Bonhomme, Lamadon, and Manresa, 2017)
- These are (currently) largely application-specific

# When should I use machine learning?

- If you are mainly interested in prediction
- If you have an intermediate step of your model estimation that requires making predictions
- If you need to compress a prohibitively large data set

# References

Athey, Susan and Guido W. Imbens. 2015. "Machine Learning Methods for Estimating Heterogeneous Causal Effects." Working paper, Stanford University.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. "Demand Estimation with Machine Learning and Model Combination." Working Paper 20955, National Bureau of Economic Research.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2017. "Discretizing Unobserved Heterogeneity." Working paper, University of Chicago.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2):3–28.