Machine Learning and Economics: An Introduction

SUSAN ATHEY (STANFORD GSB)

Two Types of Machine Learning

SUPERVISED

Independent observations

Stable environment

Regression/prediction:

- E[Y|X=x]

Classification

- Pr(Y=y|X=x)

UNSUPERVISED

Collections of units characterized by features

- Images
- Documents
- Individual internet activity history

Find groups of similar items

Little difference across disciplines in how these methods are used

Classification

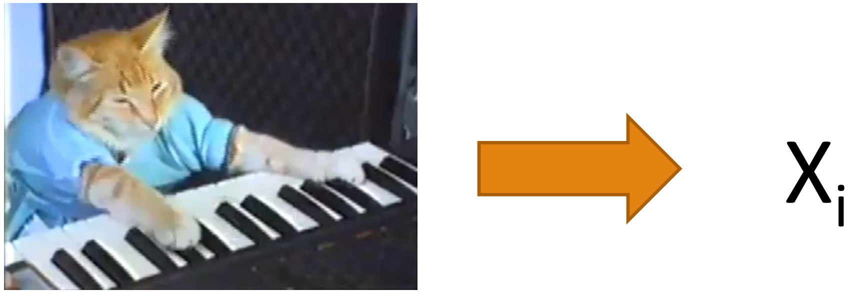
Advances in ML dramatically improve quality of image classification

Classification

Neural nets figure out what features of image are important

Features can be used to classify images

Relies on stability



X_i

Given X_i, is this a cat?

$$\Pr(Y_i = CAT|X_i) = .95$$
$$\Pr(Y_i = DOG|X_i) = .05$$

What’s New About ML?

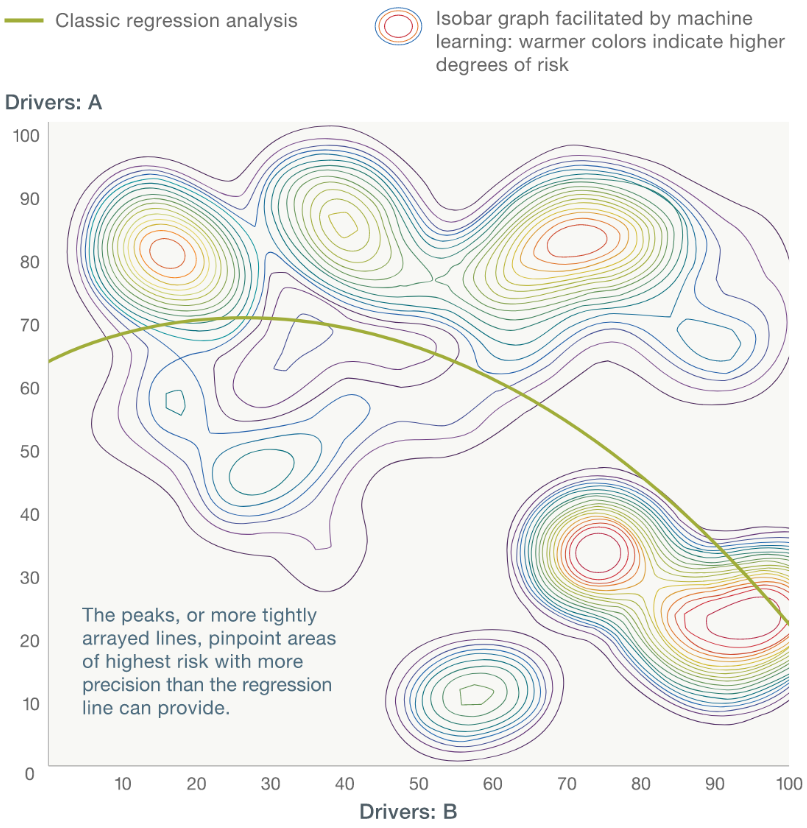
Flexible, rich, data-driven models

Increase in personalization and precision

Methods to avoid over-fitting

The contrast between routine statistical analysis and data generated by machine learning can be quite stark.

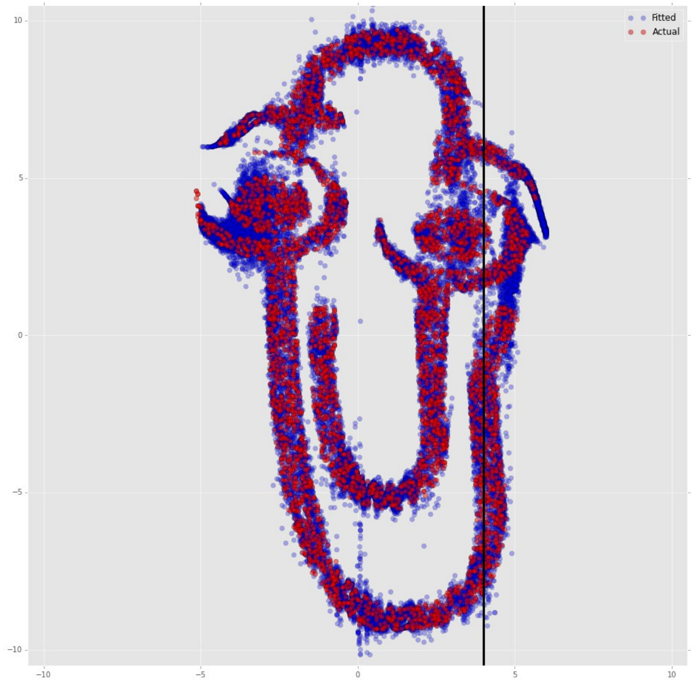
Value at risk from customer churn, telecom example

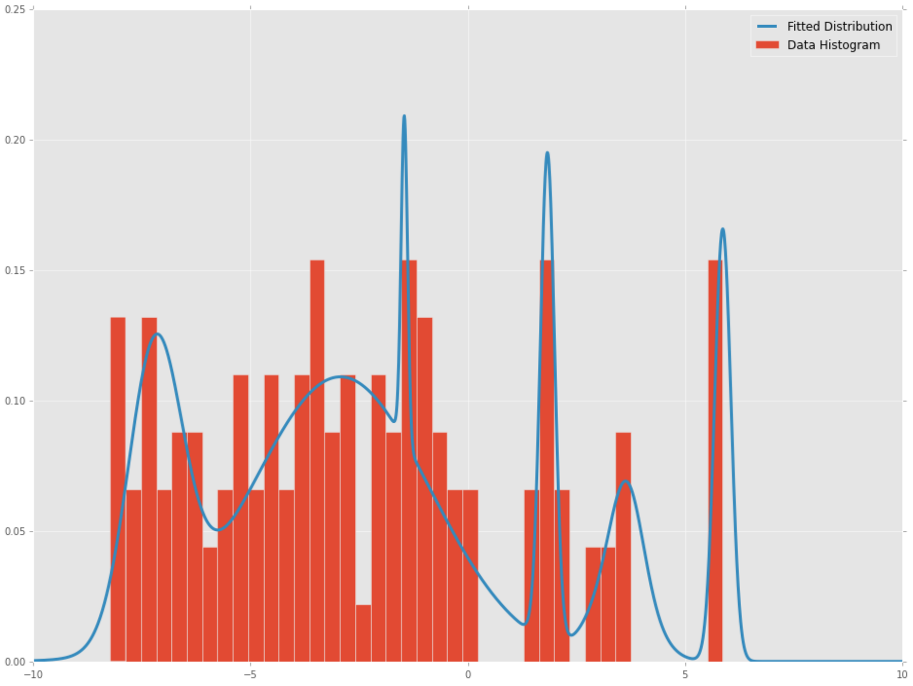


The peaks, or more tightly arrayed lines, pinpoint areas of highest risk with more precision than the regression line can provide.

McKinsey&Company

Ability to Fit Complex Shapes





Prediction in a Stable Environment

Goal: estimate $\mu(x) = E[Y|X = x]$ and minimize MSE in a new dataset where only *X* is observed

- MSE: $\frac{1}{I}\sum_i \left(Y_i - \hat{\mu}(X_i)\right)^2$
- No matter how complex the model, the output, the prediction, is a single number
- Can hold out a test set and evaluate the performance of a model
- Ground truth is observed in a test set
- Only assumptions required: independent observations, and joint distribution of (*Y*,*X*) same in test set as in training set

Note: minimizing MSE entails bias-variance tradeoff, and always accept some bias

- Idea: if estimator too sensitive to current dataset, then procedure will be variable across datasets
- Models are very rich, and overfitting is a real concern, so approaches to control overfit necessary

Idea of ML algorithms

- Consider a family of models
- Use the data to select among the models or choose tuning parameters
- Common approach: cross-validation
  - Break data into 10 folds
  - Estimate on 9/10 of data, estimate MSE on last tenth, for each of a grid of tuning parameters
  - Choose the parameters that minimize MSE

ML works well because you can accurately evaluate performance without add'l assumptions

- Your robotic research assistant then tests many models to see what performs best

What We Say v. What We Do (ML)

What we say

- ML = Data Science, statistics
  - Is there anything else?
- Use language of answering questions or solving problems, e.g. advertising allocation, salesperson prioritization
- Aesthetic: human analyst does not have to make any choices
- All that matters is prediction

What we do

- Use predictive models and ignore other considerations
  - E.g. Causality, equilibrium or feedback effects
- Wonder/worry about interpretability/reliability/robustness/ adaptability, but have little way to ask algos to optimize for it

Contrast with Traditional Econometrics

Economists have focused on the case with substantially more observations than covariates (N>>P)

- In-sample MSE is a good approximation to out-of-sample MSE
- OLS is BLUE, and if overfitting is not a problem, then no need to incur bias
- OLS uses all the data and minimizes in-sample MSE

OLS obviously fails due to overfitting when P~N and fails entirely when P>N

- ML methods generally work when P>N

Economists worry about estimating causal effects and identification

- Causal effects
- Counterfactual predictions
- Separating correlation from causality
- Standard errors
- Structural models incorporating behavioral assns

Identification problems can not be evaluated using a hold-out set

- If joint dist'n of observable same in training and test, will get the same results in both

Causal methods sacrifice goodness-of-fit to focus only on variation in data that identifies parameters of interest

What We Say v. What We Do (Econometrics)

What We Say

- Causal inference and counterfactuals
- God gave us the model
- We report estimated causal effects and appropriate standard errors
- Plus a few additional specifications for robustness

What we do

- Run OLS or IV regressions
  - Try a lot of functional forms
  - Report standard errors as if we ran only one model
  - Have research assistants run hundreds of regressions and pick a few "representative" ones
- Use complex structural models
  - Make a lot of assumptions without a great way to test them

Key Lessons for Econometrics

Many problems can be decomposed into predictive and causal parts

- Can use off-the-shelf ML for predictive parts

Data-driven model selection

- Tailored to econometric goals
  - Focus on parameters of interest
  - Define correct criterion for model
  - Use data-driven model selection where performance can be evaluated
- While retaining ability to do inference

ML-Inspired Approaches for Robustness

Validation

- ML always has a test set
- Econometrics can consider alternatives
  - Ruiz, Athey and Blei (2017) evaluate on days with unusual prices
  - Athey, Blei, Donnelly and Ruiz (2017) evaluate change in purchases before and after price changes
  - Tech firm applications have many A/B tests and algorithm changes

Other computational approaches for structural models

- Stochastic gradient descent
- Variational Inference (Bayesian models)

See Sendhil Mullainathan et al (JEP, AER) for key lessons about prediction in economics

- See also Athey (Science, 2017)

Empirical Economics in Five Years: My Predictions

Regularization/data-driven model selection will be the standard for economic models

Prediction problems better appreciated

Measurement using ML techniques an important subfield

Textual analysis standard (already many examples)

Models will explicitly distinguish causal parts and predictive parts

Reduced emphasis on sampling variation

Model robustness emphasized on equal footing with standard errors

Models with lots of latent variables

ML and Causal Inference: An Overview of Recent Work

(NOT A COMPREHENSIVE LITERATURE REVIEW)

Themes from ML/Causal Inference Literature

Off-the-shelf ML methods…

- Cannot typically be DIRECTLY used for statistical inference
- Parameter estimates are biased due to MSE optimization
- Produce biased estimates of causal effects due to confounding
- Are not optimized for the objective

But a variety of modifications and tricks can help…

- Using ML to categorize units, then analyze groups
- Sample splitting & cross-fitting (e.g., with subsampling for forests)
- Changing optimization criterion to optimize for parameter(s)
- Estimate nuisance parameters using std. ML and use orthogonalization
- Find data sets with many "mini-experiments," design test sets to validate, do careful "off-policy" evaluation

Average Treatment Effects (ATE) w/ Unconfoundedness

- Targeted ML (van der Laan et al, series)
- Double-LASSO (Belloni, Chernozhukov, Hansen)
- Residual Balancing (Athey, Imbens, Wager (2016), Hirshberg and Wager (2017))
- Double ML (Chernozhukov et al 2017)
- Methods based on averaging CATE:
  - Generalized Random Forests (Athey, Tibshirani, Wager 2016)
  - BART (Chipman and George, 2010), e.g. Hill

Conditional ATE (CATE) w/ Unconf.:

Low-Dimensional Treatment Effect Heterogeneity

“Moving the Goalposts”

- Targeted ML (van der Laan et al)
- LASSO-based methods (Imai and Ratkovic)
- Causal Trees (Athey and Imbens, PNAS 2016)
- X-Learners (Kunzel, Sekhon, Bickel, Yu, 2017)
- Chernozukov and Duflo (2018)

Conditional ATE w/ Unconf.:

Non-parametric Case

- Causal Forests (Wager and Athey, 2015)
- Generalized Random Forests (Athey, Tibshirani, Wager 2016)
- Nie and Wager (2017)

Optimal (Personalized) Policy Estimation

Offline

- From ML Literature: Strehl et al. (2010); Dudik et al. (2011); Li et al. (2012); Dudik et al. (2014); Li et al. (2014); Swaminathan and Joachims (2015); Jiang and Li (2016); Thomas and Brunskill (2016); Kallus (2017).
- ML + Semiparametric efficiency: Efficient Policy Estimation (Athey and Wager, 2021; Zhou, Athey, and Wager, 2019)

Policy Estimation Online:

Contextual Bandits

- Very large ML literature; see e.g. Li et al. (2010), Goldenshluger and Zeevi (2013), Li et al. (2017), Bastani and Bayati (2015), Feraud et al. (2016), Krishnamurthy et al (2021)
- Estimation issues (Athey et al 2018), hypothesis testing (Kasy 2019; Athey et al 2019), and more

Supplementary Analyses

- Robustness of parameter estimates (Athey, Imbens 2015)
- Confoundedness (Athey, Imbens, Pham, Wager, 2017)
- See also Athey-Imbens 2017 survey

Instrumental Variables (IV) Estimates (Average or Low-D CATE)

- Targeted ML
- LASSO-Based methods (Belloni, Chernozhukov, Hansen et al, series)

IV: Heterogeneous Effects (CLATE)

- ML/GMM Trees (Zeiles et al 2008; Athey, Tibshirani, Wager 2016; Asher et al 2016
- Generalized Random Forests (Athey, Tibshirani, Wager 2016)
- Deep IV [neural nets] (Lewis, Leyton-Brown and Taddy, 2016)

Heterogeneous Parameter Estimation in GMM/ML Models

(Non-parametric heterogeneity)

- Generalized Random Forests (Athey, Tibshirani, Wager 2016)

Regression Discontinuity

- Local Linear Forests (Athey, Friedberg, Wager 2020)

Panel Data

(Diffs-in-Diffs, Synthetic Controls, etc.)

- Synthetic Controls with Regularized Regression for weights (Doudchenko and Imbens 2016)
- Matrix Completion w/ Nuclear Norm (Athey et al 2017)

Combining Observational and Experimental Data

- Surrogates (Athey, Chetty, Imbens, Kang, 2017)
- Peysakhovich and Lada (2016)

Large-Scale Structural Models (Consumer Demand)

- Bayesian matrix factorization for independent categories (Athey, Blei, Donnelly, Ruiz, 2017; Athey, Blei, Donnelly, Ruiz, Schmidt, 2018)
- Matrix factorization for multi-step shopping decisions (Wan et al, 2017)
- Estimating complements/substitutes with many items in Bayesian model (Ruiz, Athey, Blei 2017)