Machine Learning Methods That Economists Should Know About

Athey & Imbens (2019)

Xi Xiang

NJUBS, Department of Economics

2026.3.22

Outline

Outline

Two Cultures in Data Analysis

Econometrics

- Structural models
- Target estimands
- Inference, asymptotics

Machine Learning

- Prediction-oriented
- Flexible approximations
- Out-of-sample performance

Core tension: inference vs. prediction

Why ML Was Slow to Enter Economics

- Valid inference after model selection is non-trivial
- ML often ignores causal structure
- Economic questions focus on *parameters*, not predictions

Paper's message: ML complements econometrics when embedded in economic structure.

Outline

Different Objectives

Econometrics targets a parameter:

$$\theta = \theta(P), \quad \sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, V)$$

Machine learning targets a function:

$$\min_f \ \mathbb{E}[\ell(Y, f(X))]$$

- Econometrics: interpretability + inference
- ML: predictive accuracy + scalability

Cross-Validation and Regularization

K-fold CV:

$$\mathrm{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \ell(Y_i, \hat{f}_{-k,\lambda}(X_i))$$

Regularization:

$$\hat{f} = \arg\min_{f} \sum_{i} \ell(Y_i, f(X_i)) + \lambda \cdot \mathrm{Complexity}(f)$$

- Bias–variance tradeoff
- λ chosen by data, not theory

Approximate Sparsity

High-dimensional linear model:

$$Y = X'\beta_0 + \varepsilon, \quad p \gg n$$

Assumption:

$$\sum_{j=1}^{p} |\beta_{0j}|^q \leq C, \quad 0 < q < 1$$

- Few important controls, many negligible ones
- Justifies LASSO-type methods

Penalized Estimation and Computation

LASSO:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i} (Y_i - X_i'\beta)^2 + \lambda \|\beta\|_1$$

- Convex optimization
- Feasible for massive datasets
- Trade structure for scalability

Debiased LASSO

Naive post-selection inference fails.
Debiased estimator:

$$\tilde{\beta}_j = \hat{\beta}_j + \frac{1}{n}\sum_{i=1}^{n} \hat{\Theta}_j' X_i (Y_i - X_i' \hat{\beta})$$

- Restores asymptotic normality
- Enables valid confidence intervals

Outline

Conditional Mean Estimation

Target:

$$g(x) = \mathbb{E}[Y|X = x]$$

- High-dimensional
- Nonlinear and interactive

Outline

Regularized Linear Models

$$\min_{\beta} \sum_{i} (Y_i - X_i'\beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- Ridge: shrinkage
- LASSO: sparsity
- Elastic Net: correlated regressors

Outline

Regression Trees

Partition covariate space:

$$\hat{g}(x) = \sum_{m=1}^{M} c_m \mathbf{1}(x \in R_m)$$

- Captures interactions automatically
- High variance individually

Random Forests

Ensemble of trees:

$$\hat{g}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{g}^{(b)}(x)$$

- Bootstrap aggregation
- Robust to irrelevant covariates

Neural Networks

Feedforward network:

$$f(x) = W_L \sigma(W_{L-1} \sigma(\cdots \sigma(W_1 x)))$$

- Universal approximation
- Optimization via SGD

Outline

Classification Problem

Outcome $Y \in \{0, 1, \ldots, K\}$. Decision rule $f : X \to \mathcal{Y}$ minimizes risk:

$$\min_f \ \mathbb{E}[\ell(Y, f(X))]$$

- 0–1 loss and surrogate losses
- Focus on decision boundaries, not conditional means

Outline

Classification Trees

Partition feature space into regions $\{R_m\}$:

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \mathbf{1}(x \in R_m)$$

- Greedy recursive splitting
- Interpretable but high variance

Random Forests for Classification

Ensemble classifier:

$$\hat{f}_{RF}(x) = \text{majority vote}\{\hat{f}^{(b)}(x)\}_{b=1}^{B}$$

- Variance reduction via bagging
- Robust to overfitting

Outline

Support Vector Machine: Primal Form

Linear SVM solves:

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$Y_i(w'X_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

- Maximizes margin
- Allows misclassification via slack variables

SVM: Dual Formulation

Dual problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j X_i' X_j$$

subject to:

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i Y_i = 0$$

- Depends only on inner products $X_i' X_j$
- Sparse solution: many $\alpha_i = 0$

Support Vectors

Only observations with $\alpha_i > 0$ matter.
Decision rule:

$$f(x) = \sum_{i:\alpha_i>0} \alpha_i Y_i X_i' x + b$$

- Margin determined by few critical points
- Robust to irrelevant observations

Kernel Trick

In the dual:

$$X_i' X_j \;\Rightarrow\; K(X_i, X_j)$$

No need to compute feature map explicitly.
Common kernels:

- Polynomial:

$$K(x, z) = (1 + x'z)^d$$

- Gaussian (RBF):

$$K(x, z) = \exp(-\|x - z\|^2 / 2\sigma^2)$$

Nonlinear Decision Boundaries

Decision function:

$$f(x) = \sum_i \alpha_i Y_i K(X_i, x) + b$$

- Linear in feature space
- Nonlinear in original covariates
- Effective in high dimensions

Outline

Clustering

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - \mu_k\|^2$$

- Market segmentation
- Firm and worker typologies

Dimension Reduction

Principal components:

$$\max_{v} \mathbb{V}ar(Xv) \text{ s.t. } \|v\| = 1$$

- Summarize information
- Often used as first-stage controls

Outline

Outline

Treatment Effects and Targets

Potential outcomes:

$$Y(1),\ Y(0)$$

Average Treatment Effect (ATE):

$$\theta = \mathbb{E}[Y(1) - Y(0)]$$

Conditional Average Treatment Effect (CATE):

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Goal: estimate causal effects, not predictive accuracy.

Identification Assumptions

Observed data:

$$W = (Y, D, X)$$

Key assumptions:

- **Unconfoundedness**:

$$(Y(1), Y(0)) \perp D \mid X$$

- **Overlap**:

$$0 < P(D = 1 \mid X) < 1$$

Allows identification of causal effects from observables.

Nuisance Functions

Two key high-dimensional objects:
Outcome regression:

$$m_d(x) = \mathbb{E}[Y \mid D = d, X = x], \quad d \in \{0, 1\}$$

Propensity score:

$$e(x) = P(D = 1 \mid X = x)$$

Both can be complex, nonlinear, and high-dimensional.

Why Naive Plug-in Fails

Plug-in estimators:

$$\hat{\theta} = \frac{1}{n} \sum_i \left(\hat{m}_1(X_i) - \hat{m}_0(X_i) \right)$$

Problems:

- ML estimators converge slowly
- Regularization bias contaminates $\hat{\theta}$
- Standard asymptotics fail

Outline

Orthogonal (Neyman) Scores

Define score function:

$$\psi(W;\theta,\eta) = (D - e(X))(Y - m(X)) + \theta$$

where

$$\eta = (m(\cdot), e(\cdot))$$

Key property:

$$\frac{\partial}{\partial \eta}\mathbb{E}[\psi(W;\theta,\eta)]\Big|_{\eta=\eta_0} = 0$$

Double Machine Learning (DML)

Procedure:

1. Use ML to estimate $m(X)$ and $e(X)$
2. Apply sample splitting / cross-fitting
3. Plug estimates into orthogonal score

Result:

$$\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, V)$$

- Valid inference with flexible ML
- Bias from first stage is second-order

Outline

Estimating Heterogeneous Effects

Beyond ATE: estimate $\tau(x)$.
Methods:

- Causal Trees
- Causal Forests
- Honest sample splitting

Goal: uncover systematic treatment effect heterogeneity.

Outline

From Static Experiments to Adaptive Design

Classical randomized experiments:

- Fixed treatment assignment
- Focus on unbiased estimation

Limitations:

- Inefficient learning
- Ethical or cost concerns

ML enables **adaptive experimentation**.

Multi-Armed Bandit Problem

At each time t:

- Choose action $A_t \in \{1, \ldots, K\}$
- Observe reward $Y_t(A_t)$

Objective:

$$\max \mathbb{E}\Big[\sum_{t=1}^{T} Y_t(A_t)\Big]$$

Key tradeoff:

- Exploration vs exploitation

Regret and Learning

Regret:

$$\text{Regret}(T) = \sum_{t=1}^{T} \left(Y_t(A^*) - Y_t(A_t) \right)$$

Algorithms:

- ε-greedy
- Upper Confidence Bound (UCB)
- Thompson Sampling

Goal: sublinear regret.

Reinforcement Learning

Markov Decision Process (MDP):

- State S_t
- Action A_t
- Reward R_t
- Transition $P(S_{t+1} \mid S_t, A_t)$

Objective:

$$\max_{\pi} \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t R_t\Big]$$

Economic Applications

- Dynamic pricing
- Online advertising
- Policy learning and experimentation
- Education and labor market programs

RL = dynamic treatment assignment with learning.

Outline

Motivation

Observed data:

$$Y_{ij} = \text{outcome of user } i \text{ on item } j$$

Most entries missing.

Examples:

- Consumer-product choices
- Firm-worker matches
- Student-course evaluations

Low-Rank Structure

Matrix form:

$$Y \in \mathbb{R}^{N \times J}$$

Assumption:

$$\text{rank}(Y) \ll \min(N, J)$$

Interpretation:

- Few latent preferences
- Few latent product characteristics

Matrix Completion

Optimization problem:

$$\min_{M} \sum_{(i,j)\in\Omega} (Y_{ij} - M_{ij})^2 \quad \text{s.t. } \operatorname{rank}(M) \leq r$$

Convex relaxation:

$$\min_{M} \sum_{(i,j)\in\Omega} (Y_{ij} - M_{ij})^2 + \lambda \|M\|_*$$

$\|\cdot\|_*$: nuclear norm

Recommendation Systems

Prediction:

$$\hat{Y}_{ij} = u_i' v_j$$

- Personalized recommendations
- Matching markets
- Demand estimation with sparse data

Links ML to discrete choice and factor models.

Outline

Text as Data

Text sources:

- Policy documents
- Earnings calls
- Contracts and regulations

Challenge:

- High-dimensional
- Unstructured

Text Representation

Bag-of-words:

$$X_{id} = \text{frequency of word } d$$

TF-IDF weighting.

Topic models:

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

Documents as mixtures of latent topics.

Text in Econometric Analysis

Uses:

- Controls for latent information
- Measurement of beliefs and expectations
- Policy evaluation

ML tools:

- LASSO on text features
- Embeddings + DML

Challenges

- Interpretability
- Measurement error
- Text is endogenous

Economic theory still essential.

Outline

Conclusion: Machine Learning for Economics

This paper provides a unified framework for integrating machine learning into economic analysis.

- **Conceptual distinction**: Prediction and causal inference are fundamentally different goals.
- **Methodological contribution**: ML tools (regularization, trees, kernels, neural networks) are valuable for estimating high-dimensional objects, not for replacing economic structure.
- **Inference with ML**: Approximate sparsity, orthogonal scores, and sample splitting enable valid inference despite data-driven model selection.
- **Causal machine learning**: Outcome regressions and propensity scores can be flexibly estimated while preserving identification and asymptotic guarantees.
- **Beyond static models**: ML extends naturally to adaptive experiments, dynamic decision problems, matrix completion, recommendation systems, and text data.

What Does ML Change for Economists?

Machine learning changes *how* economists work, not *what* economists study.

- Allows rich, high-dimensional controls without ad hoc modeling
- Shifts focus from model selection to target parameters
- Enables data-driven discovery under formal identification
- Expands empirical scope to new data types and decision problems

Open challenges:

- Interpretability vs flexibility
- Dynamic and strategic environments
- Unified theory of inference for complex ML systems