Chapter 12
Machine Learning and Model Selection: Outline

Agoston Reguly, Esfandiar Maasoumi and László Mátyás

Table of Contents

1. Brief history of Machine Learning methods
2. Brief history of Model Selection in Econometrics
3. The meeting of the two approaches
   a. Large enough samples and cross-validation
4. Forgotten lessons from model selections and why machine learning seems to win
   a. The (strict) endogeneity assumption
   b. Prediction exercise vs causality
   c. Relevance of the asymptotic behaviour
5. When parameters matter → causal ML
   a. Breakthroughs in the 2010s
6. Why ML is not the benchmark in current empirical analysis
   a. (In)stability of the method
   b. Better use as an exploratory tool
   c. Challenges of interpretation - still black-box models, but now with asymptotics
7. ML methods that did not came through: data mining and unsupervised ML
8. Will LLMs ever matter for econometrics?

Agoston Reguly ✉
Corvinus University of Budapest, Budapest, Hungary and Georgia Institute of Technology, Atlanta, Georgia, USA, e-mail: agoston.reguly@uni-corvinus.hu

Esfandiar Maasoumi
Emory University, Atlanta, Georgia, USA, e-mail: emaasou@emory.edu

László Mátyás
Central European University, Budapest, Hungary, and Vienna, Austria, e-mail: matyas@ceu.edu

12.1 Brief History of Machine Learning up to 2000s

- **1940s–1950s: Foundations**
  - 1943: McCulloch & Pitts propose the first mathematical model of a neuron.
  - 1950: Alan Turing introduces the Turing Test.
  - 1957: Rosenblatt develops the Perceptron.
- **1960s–1970s: Early Algorithms**
  - 1960s: Emergence of nearest neighbor and linear regression methods.
  - 1967: Nearest Neighbor algorithm introduced.
  - 1970s: Development of backpropagation theory.
  - First 'AI Winter' due to limited computing power.
- **1980s: Revival and Neural Networks**
  - 1980s: Machine learning becomes distinct from AI.
  - 1986: Rumelhart, Hinton & Williams popularize backpropagation.
  - Decision trees introduced (ID3 algorithm by Quinlan).
  - 1989: Yann LeCun applies backpropagation to handwritten digit recognition.
- **1990s: Statistical Learning Era**
  - Shift to probabilistic models and data-driven approaches.
  - 1992: Support Vector Machines (SVM) introduced by Vapnik.
  - Ensemble methods: bagging (1994), boosting (1996).
  - Bayesian networks and Hidden Markov Models widely used.
  - (1998: LeCun develops LeNet-5, CNN for digit recognition).
- **2000s: Big Data and Kernel Methods**
  - Machine learning central to data mining and web applications.
  - Kernel methods and SVMs dominate.
  - Growth of unsupervised learning and reinforcement learning.
  - Focus on scalability and real-world applications.

12.2 Brief History of Model Selection in Econometrics up to 2000

- **1950s–1960s: Early Specification and Theory-Driven Models**
  - Model selection was primarily theory-driven, relying on economic theory for specification.
  - Hypothesis testing (t-tests, F-tests) used for variable inclusion/exclusion.
  - Late 1960s: Box & Jenkins introduce ARIMA models, marking a shift toward data-driven forecasting.
  - Pursuing R^2 to select "best models"

- **1970s: Rise of Information Criteria and Specification Analysis**
  - 1973: Akaike introduces AIC (Akaike Information Criterion).
  - 1978: Schwarz introduces BIC (Bayesian Information Criterion).
  - Mallows' Cp (1973) emerges for regression model selection.
  - Leamer (1983) warns against specification searches and advocates loss functions and priors.
- **1980s: Formalization and Critiques**
  - Handbook of Econometrics consolidates statistical theories of model selection.
  - Stepwise regression, cross-validation, and goodness-of-fit tests become common.
  - Bayesian approaches gain traction for incorporating prior beliefs.
- **1990s: General-to-Specific (Gets) and Automated Strategies**
  - Hendry and LSE approach promote general-to-specific modeling.
  - Development of PcGive and OxMetrics enables semi-automated model selection.
  - $(EC)^2$ conferences focus on model selection and evaluation.
  - AIC and BIC become standard tools in time-series and panel data models.
- **By 2000: Consolidation of Approaches**
  - Model selection combines theory-consistent and data-admissible strategies.
  - Common tools: AIC, BIC, Mallows' Cp, cross-validation, encompassing tests, Bayesian methods.
  - Persistent challenges: balancing complexity vs. parsimony, avoiding overfitting, addressing model uncertainty.

12.3 Cross-validation

- **Definition**: Cross-validation (CV) is a resampling technique used to assess the predictive performance of a model by splitting data into training and validation sets multiple times.
- **Purpose in Econometrics**: Traditionally, econometrics relies on information criteria (AIC, BIC) for model selection. CV introduces a data-driven approach that focuses on out-of-sample prediction accuracy. It needs larger dataset, as it relies on splitting the data
- **Types of CV**:
  - *k-Fold CV*: Data is divided into k subsets; each fold is used once for validation.
  - *Leave-One-Out CV (LOOCV)*: Each observation acts as a validation set; useful for small samples.
  - *Time-Series CV*: Rolling or expanding windows to respect temporal dependence.

- **Advantages Over AIC/BIC**:
  - CV does not assume a correct model specification.
  - Better suited for high-dimensional settings and machine learning algorithms (e.g., LASSO, Ridge).
- **Applications in Econometrics**:
  - Regularization parameter tuning (e.g., choosing λ in LASSO).
  - Forecast evaluation in macroeconomic and financial models.
  - Model comparison when theoretical guidance is weak.
- **Challenges**:
  - Computational cost for large datasets.
  - Dependence structures (e.g., autocorrelation in time series) require adapted CV methods.
- **Integration with ML**:
  - CV is central to Machine Learning frameworks, especially for methods, which uses causal inference.
  - Used also for ensemble methods and nonparametric econometric models.

12.4 Forgotten Lessons from Model Selection and Why Machine Learning Seems to Win

- **Strict Endogeneity Assumption**:
  - Classical econometric models assume regressors are strictly exogenous (uncorrelated with the error terms).
  - In practice, this assumption is often violated, especially in observational data.
  - Machine learning methods prioritize prediction and can handle correlated features without requiring strict exogeneity, though at the cost of causal interpretation.
- **Prediction vs. Causality**:
  - Econometrics traditionally emphasizes causal inference, requiring identification strategies and valid instruments.
  - Machine learning focuses on minimizing prediction error, often ignoring causal structure.
  - This shift means ML "wins" in predictive tasks but may fail in policy analysis without additional econometric tools.
- **Relevance of Asymptotic Behavior**:
  - Econometric theory relies heavily on asymptotic properties (consistency, efficiency) under large samples.

  - Machine learning methods often prioritize finite-sample performance and generalization error, using cross-validation rather than asymptotic guarantees.
  - In high-dimensional settings, asymptotic results may be less informative, making ML approaches more practical.

- **Why Machine Learning Seems to Win**:
  - Flexibility in high-dimensional data: ML algorithms (e.g., LASSO, Random Forests) adapt well to large feature spaces without strong parametric assumptions.
  - Robustness to misspecification: Econometric models assume correct functional form; ML methods are more robust through nonparametric or ensemble approaches.
  - Data-driven model selection: ML uses cross-validation and regularization rather than theory-driven selection, improving predictive accuracy.
  - Computational advances: ML benefits from scalable algorithms and computing power, making complex models feasible in practice.

12.5 Causal Machine Learning and breakthroughs in Asymptotic Behavior

- **Causal Machine Learning**:
  - *Goal*: Extend machine learning beyond prediction to estimate causal effects (e.g., treatment effects, policy impacts).
  - *Key Methods*:
    - **Double Machine Learning (DML)**: Combines ML for nuisance parameter estimation with econometric orthogonalization for unbiased causal inference.
    - **Generalized Random Forests**: Estimate heterogeneous treatment effects using tree-based methods.
  - *Why It Matters for Econometrics*: Economists need interpretable, theory-consistent estimates for policy decisions. ML provides flexibility in high-dimensional settings but must respect identification conditions.
  - *Bridging Econometrics and ML*: Econometricians demand theoretical rigor; asymptotic proofs make ML methods acceptable for causal analysis. This is done in the last decade.

12.6 Why Machine Learning is Not Yet the Benchmark in Current Econometric Analysis

- **Instability of Methods and Results**:
  - Many ML algorithms are sensitive to hyperparameter choices and data splits.

  - Small changes in the dataset can lead to significantly different models and predictions.
  - This instability raises concerns for policy analysis and reproducibility.

- **Better Suited as an Exploratory Tool**:
  - ML excels at variable selection and discovering functional forms in high-dimensional data.
  - However, it often lacks the theoretical structure needed for causal inference and hypothesis testing.
  - Econometric models remain essential for formal inference and policy evaluation.

- **Challenges in Interpretation**:
  - ML models are often "black boxes" with limited interpretability.
  - Although "interpretable" ML is gaining attention, the final messages are often complicated.
  - Economists require clear parameter estimates and economic meaning, which ML does not always provide.
  - Lack of transparency complicates communication of results to fellow researchers, but also to policymakers and stakeholders.

12.7 Why Data Mining and Unsupervised Machine Learning Are Not Successful in Econometrics

- **Focus on Causality, Not Clustering**:
  - Econometrics is primarily concerned with causal inference and hypothesis testing.
  - Unsupervised ML focuses on pattern discovery (e.g., clustering, dimensionality reduction) without causal interpretation.

- **Lack of Economic Theory Integration**:
  - Unsupervised methods often ignore theoretical constraints and identification strategies central to economic analysis.
  - Economists require models that align with structural assumptions, which clustering or association rules rarely provide.

- **Interpretability Challenges**:
  - Results from unsupervised learning (e.g., clusters, latent factors) are hard to map to meaningful economic concepts.
  - Policymakers and researchers need clear, interpretable parameters rather than opaque groupings.

- **Data Structure Limitations**:

  - Economic data often have strong temporal, panel, or hierarchical structures that unsupervised ML does not handle well.
  - Methods like PCA or k-means assume independence, which is unrealistic in economic contexts.

- **Risk of Spurious Patterns**:
  - Data mining can identify correlations without theoretical justification, leading to misleading conclusions.
  - Econometrics emphasizes avoiding spurious relationships through identification and robustness checks.

- **Limited Use Cases for Policy Analysis**:
  - While clustering might help segment markets or consumers, it rarely provides actionable causal insights for policy evaluation.

12.8 Open Questions About the Use of Large Language Models (LLMs) in Econometrics

- **Interpretability and Economic Meaning**:
  - How can LLM outputs be mapped to interpretable economic concepts?
  - LLM strength is allowing to use many unstructured data (e.g. text) and convert it to measures (variables), which allows specific economic analysis. But can we trust this transition? Is it consistent or not?

- **Data Privacy and Confidentiality**:
  - Economic datasets often contain sensitive information.
  - How can LLMs be trained or fine-tuned without violating privacy regulations?

- **Bias and Fairness**:
  - LLMs inherit biases from training data.
  - How do these biases affect policy recommendations or economic forecasts?

12.9 Conclusion