Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 1: Introduction (ISL ch. 1)

Instructor: Ma, Jun

Renmin University of China

September 9, 2021

What is machine learning?

- Literally, Machine learning (ML) means that the machine (computer) learns and produces meaningful results after data comes in.
- Machine learning (ML) is a flourishing field that encompasses many methods for data-driven prediction in different contexts.
- Thomas Sargent's (Nobel Laureate in Economics) speech in 2018 World Forum on Scientific and Technological Innovation:

  > *"Artificial intelligence is, first of all, gorgeous rhetoric. Artificial intelligence is actually statistics, but with a very gorgeous phrase, in fact, is statistics."*

- Many tools that are considered as ML methods indeed have a long history in the statistical literature:
  - Regularization/penalty: Akaike (1974)
  - Linear discriminant analysis: Fisher (1936)
  - Regression trees: Morgan and Sonquist (1963)
  - Deep learning and neural networks (mid 20th century)

Real-life applications of ML methods

- ML successfully expanded its audience. Nowadays it is of interest to industrial practitioners, while it was of primarily academic interest decades ago.
- Current explosion of interest is due to digital economy, data abundance and contribution from computer scientists that overcomes computational difficulty.
- Examples of industrial applications:
  - Industrial practitioners are mostly interested in out-of-sample prediction.
  - Product recommendation: using a customer's purchase history to recommend new products.
  - Banking: identify customers most likely to default.
  - Finance: automatic trading.
  - Spam email detection.
  - Image recognition.

Types of ML

- Supervised:
  - Data on the outcome variable and predictors (regressors).
  - Train (estimate) the model to predict the outcomes from regressors.
  - An example: get a sample of pictures of cats and dogs labeled as “cat” or “dog”, train a model to classify an image as “cat” or “dog” and classify an unseen image as “cat” or “dog”.
- Unsupervised:
  - Only data on predictors, no outcome variable.
  - Explore data to reveal the data structure (clusters or groups).
  - An example: get a sample of unlabeled pictures of cats and dogs and ask the computer to classify the pictures into two groups.

Statistics of Big Data

- From statistical point of view, most ML methods are designed for analysis of big data.
- Big data: number of regressors/predictors ≈ (or even >) number of observations.
  - 2000 genetic characteristics (predictors) of tumor types, 60 observations.
  - Data scraping for political economy: thousands of words, count how often they appear in articles online.
- Classical methods (e.g., classical multiple linear regression or classical nonparametric regression) can not be applied to big data.
- Theory does not always predict which of many highly correlated predictors should be included. LASSO provides data-driven model selection for prediction and inference.

Linear regression with big data

- Linear regression model with n observations and p regressors (big data: $p \approx n$ or even $p > n$):

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i} + U_i,\ i = 1, 2, ..., n.$$

- The OLS estimator is not uniquely defined.
- Perfect multicollinearity: The sample of dependent variables can be perfectly explained by the regressors.
- Perfect in-sample fit ($R^2 = 1$). But we are interested in out-of-sample prediction, which can be of poor accuracy.
- Overfitting: capturing relationships that only appear in the sample so that the accuracy of out-of-sample prediction is much worse than that of in-sample prediction.
- Statistically, overfitting happens due to large variance. LASSO uses regularization to reduce the variance.

Least Absolute Shrinkage and Selection Operator (LASSO)

- The LASSO is just constrained least squares: for some $c > 0$,

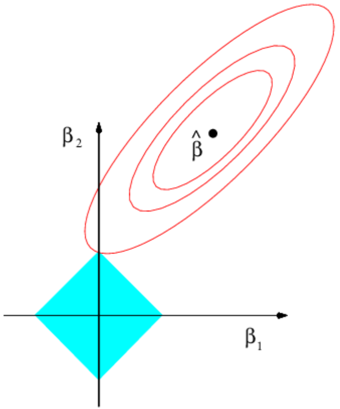
$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{1,i} - \cdots - \beta_p X_{p,i}\right)^2 \text{ s.t. } \sum_{j=1}^{p} \left|\beta_j\right| \leq c.$$

- By the Lagrangian multiplier method, LASSO can be viewed as regularized least squares:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{1,i} - \cdots - \beta_p X_{p,i}\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|.$$

  $\lambda > 0$ is a tuning parameter. We let the data tell us what λ should be (data-driven selection of the tuning parameter).

- Most ML methods are statistical methods with data-driven regularization, which often needs sophisticated algorithms.



- $\hat{\beta}$: the OLS estimator.
- Red contours: constant residual sum of squares with respect to (β_1, β_2).
- Shaded rectangle: $\{(\beta_1, \beta_2) : |\beta_1| + |\beta_2| \leq c\}$.
- LASSO forces some of regression coefficients to be zero: model selection and parameter estimation in one step.

Applications of ML in econometrics

- Economists are mostly interested in causal inference, instead of prediction.
- ML tools are viewed by some leading econometricians as new-generational nonparametric methods for removing nuisance functions in econometric models for causal inference.
- ML tools are also effective and elegant in the context of big data (e.g., many controls and many instruments), when classical econometric methods do not work well.

Causality

- Natural sciences use controlled lab experiments. Experiment are often impossible in economics (too costly and/or for ethical reasons).
- Econometrics encompasses a wide range of statistical tools that allow us to estimate causal effects using observational data, which is more challenging.
- In order to say that one variable has a causal effect on another, other factors affecting the outcome must be held fixed (controlled for). If the outcome changes as the variable changes with other factors held constant, we say that the variable has a causal effect.
- The causal effect is individual-specific and unobserved. E.g., the causal effect of schooling on wages for an individual worker is the difference in wages he/she would receive if we could change his/her level of education holding all other factors constant. The counterfactual wage under a different level of education is unobserved.

Correlation is not causation

- While we are interested in causal relations, statistics allows us to establish correlations (associations) in the data.
- “Dog owners are much happier than cat owners” (reported in *Washington Post*, Apr. 5, 2019)
  - The correlation between reported happiness and dog ownership not hard to believe.
  - Is there a causal effect? In other words, letting everybody own a dog makes the whole population happier?
- Going from correlations to causation requires making untestable assumptions on the structural model that generates the data.
- ML does not understand endogeneity or causality.

Structural models in econometrics

- Suppose Y is an economic outcome variable of interest (e.g., wage rate of individual workers, academic achievement of individual students, rate of return of some asset...), X is a vector of observed explanatory variables.
- There are factors in a vector ϵ that affect the outcome and are unobserved to the researcher.
- The fact that (X, ϵ) determines Y can be formulated as a functional relationship $Y = g(X, \epsilon)$. The causal effect of some variable in X on Y is given by the partial derivative of g with respect to that variable.
- This structural model (the relation g and the distribution of (X, ϵ)) characterizes the data generating mechanism of Y. We observe a sample $\{Y_i, X_i\}_{i=1}^n$ from the model, i.e., for some unobserved ϵ_i, $Y_i = g(X_i, \epsilon_i)$.
- We wish to recover the structural relation g, but there is no hope if we do not put any restriction on the model.

- We often use economic theory to justify the assumptions: what variables are in (X, ϵ) and what is the form of g.
- Two approaches:
    - Structural approach: an economic model (an agent maximizing utility subject to constraints) provides a list of variables (X, ϵ), specifies how (X, ϵ) determines Y and the researcher chooses specific functional forms for the model's components (e.g., consumers' utility function or firms' cost function). This approach is usually more difficult to implement.
    - Non-structural (statistical) approach: the restriction on g originates from statistical concerns rather than an economic model and the list of variables (X, ϵ) comes from understanding of the decision process that determines Y and background knowledge. E.g., specify a linear model $g(X, \epsilon) = \alpha + \beta X + \epsilon$ with unknown (α, β) which can be estimated by least squares.

Examples of linear models with endogeneity

- In the linear model $Y = \alpha + \beta X + \epsilon$, recovering (α, β) (in other words, consistently estimate (α, β)) requires more restrictions put on the correlation between X and ϵ.
- Education:

$$\log(\text{Wage}) = \alpha + \beta \times \text{Years of Schooling} + \epsilon,$$

ϵ = other factors, for example, ability. Since it is very hard to control for ability, one can overestimate the return to education by relying on usual correlations.

- Size of the police force and crime:

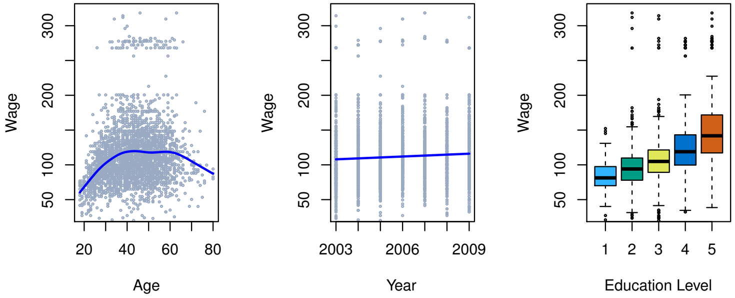
$$\text{Number of Crimes} = \alpha + \beta \times \text{Size of the Police Force} + \epsilon.$$

  Usually, cities with a lot of criminal activity have a bigger police force. Simple correlations can spuriously indicate that the size of the police force has a positive effect on the crime rates. This is an example of simultaneous equations model.
- If (X, ϵ) are correlated, an instrumental variable is needed.

Linear models with big data and endogeneity

- A data-science-based company Cinelytic uses historic data on box office performance for movies to predict box office given actors, themes, plots.
- BO_i: box office of the i-th movie; $A_{j,i}$: appearance of the j-th actor (an 0-1 variable, $A_{j,i} = 1$ if actor j appears in movie i).
- A structural model: $\mathrm{BO}_i = g\left(A_{1,i}, A_{2,i}, ..., A_{p,i}\right) + \epsilon_i$, $i = 1, 2, ..., n$, (many variables and many observations, there can be other regressors such as movie style).
- We want to know the effect of switching $A_{j,i}$ from zero to one on the box office.
- ϵ_i: unobserved factors (e.g., movie appeal to the audience). $\left(A_{1,i}, A_{2,i}, ..., A_{p,i}\right)$ are correlated with ϵ_i.
- Naive ML methods do not address the endogeneity issue.

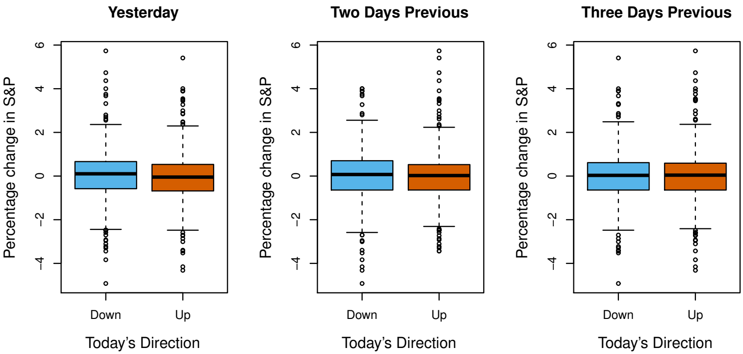
Example: the wage data



ISL Figure 1.1

- We wish to understand the association between an employee's age and education, as well as the calendar year, on his wage.
- Each variable alone is unlikely to provide an accurate prediction.
- The most accurate prediction of a given man's wage will be obtained by combining his age, his education, and the year.

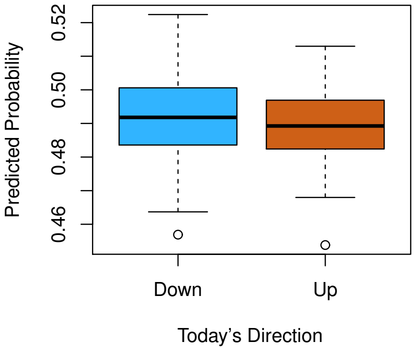
Example: the stock market data



ISL Figure 1.2

- The Wage data involves predicting a continuous or quantitative output value (regression problem).
- We may instead wish to predict a non-numerical (categorical or qualitative) value (classification problem).

- We examine a stock market data set (the Smarket data) that contains the daily movements in the Standard & Poor’s 500 (S&P) stock index over a 5-year period between 2001 and 2005.
- The goal is to predict whether the index will increase or decrease on a given day, using the past 5 days’ percentage changes in the index.
- It involves predicting whether a given day’s stock market performance will fall into the Up bucket or the Down bucket.
- The boxplots indicate little association between past and present returns.



ISL Figure 1.3

- In Chapter 4, we use ML methods to the subset data (2001–2004) and predict the probability of a decrease using the 2005 data.
- On average, the predicted probability of decrease is higher for the days in which the market does decrease. It is possible to correctly predict the direction approximately 60% of the time.