Machine Learning Tools for RCTs

1) Introduction + The Lasso

Bruno Fava

November 13, 2023

What is Machine Learning about?

- Typically, suited for prediction problems: predict Y given X
  - $\hookrightarrow$ Ex: image classification, self-driving cars, predicting treatment effects from covariates!
- Different from inference problems
  - Ex: is the (causal) population parameter θ positive?
- However, many recent developments on ML for inference!

First, let's explore the bias-variance tradeoff

- Consider the problem of estimating any θ
- Denote $\widehat{\theta}$ any estimator
- We can always decompose the Mean Squared Error (MSE):

$$\mathbb{E}\left[\left(\widehat{\theta}-\theta\right)^2\right] = \underbrace{\left(\mathbb{E}\left[\widehat{\theta}\right]-\theta\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]\right)^2\right]}_{\text{Variance}}$$

- Typical inference problem: reduce variance among unbiased estimators
  $\hookrightarrow$ Remember: OLS is BLUE!
- **However... biased estimators can have lower MSE!**

So... what is Machine Learning?

One answer...

- There is no dictionary definition of what is and isn't "Machine Learning"
- One answer...ML has two features:

  1. Regularization for bias/variance tradeoff
     $\hookrightarrow$ Allows for bias to minimize MSE
  2. Data-driven procedure for choosing amount of regularization
     $\hookrightarrow$ For example, cross-validation

- Today: example from Lasso!

Where Can Machine Learning be Useful?

- Most obviously: prediction
  - $\hookrightarrow$ For example, predicting individual TEs or group TEs
- High-dimensional estimation (under conditions)
- Variable selection
- Exploring heterogeneous treatment effects

The Lasso

LASSO - Least Absolute Shrinkage and Selection Operator

- Remember definition of OLS:

$$\widehat{\beta} = \arg\min_{b} \sum_{i=1}^{n} (Y_i - X_i' b)^2$$

- Lasso solves:

$$\widehat{\beta} = \arg\min_{b} \sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \sum_{j=1}^{p} |b_j|$$

$\hookrightarrow$ $\lambda = 0 \implies$ OLS
$\hookrightarrow$ $\lambda = \infty \implies \hat{\beta}_j = 0$ for all j

- Crucially, Lasso depends on hyperparameter $\lambda > 0$

- Intermediary values lead to some $\hat{\beta}_j = 0$, and others different from zero!

Recurring Theme in Machine Learning

- An important insight is that in-sample error is **different** from out-of-sample!

- Even though OLS minimizes sum of square error, that is done in-sample!

- But: what happens if regression has many coefficients that are truly zero?

- OLS will never set $\widehat{\beta}_j = 0$, so lots of noise!

- This could lead to poor out-of-sample predictive performance:
  - Less degrees of freedom from learning non-zero coefficients
  - More variance from including many insignificant covariates

- Lasso can help here!

  $\hookrightarrow$ Allows for bias to reduce variance and improve MSE

Beyond the Linear Model

- Note that any smooth function $f(x)$ can be approximated by a polynomial

$$f(x) \approx a_0 + a_1 x^1 + a_2 x^2 + \cdots + a_k x^k$$

$\hookrightarrow$ Approximation error made arbitrarily small by choosing large k

- With many covariates: add both powers and interactions

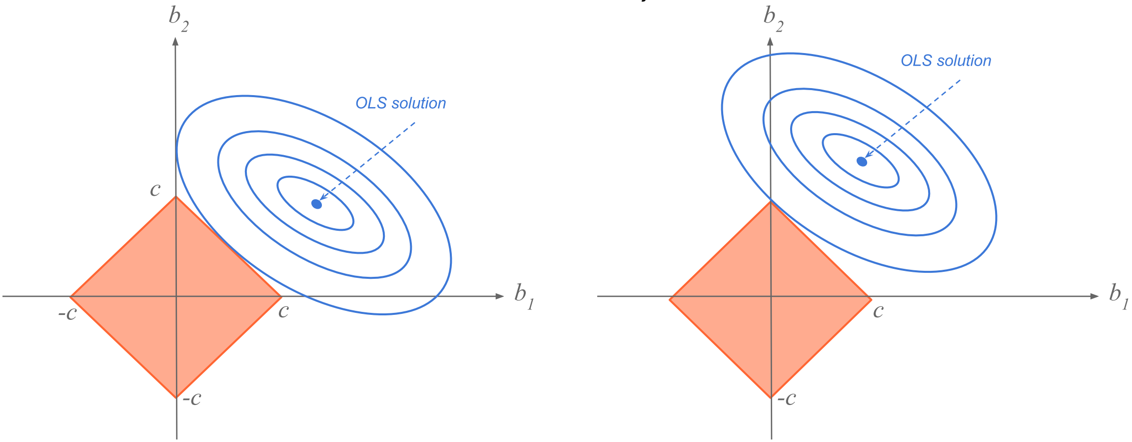
$$f(x, y) \approx a_{0,0} + a_{1,0} x^1 + a_{0,1} y^1 + a_{1,1} x^1 y^1 + a_{2,1} x^2 y^1 + \ldots$$

- Since Lasso allows for many covariates, suggests including interactions to better approximate $\mathbb{E}[Y|X]$

Why Lasso Sets Some $\widehat{\beta}_j = 0$

Note Lasso is equivalent to

$$\min_b \sum_{i=1}^{n} (Y_i - X_i' b)^2 \ \text{ s.t. } \ \sum_{j=1}^{p} |b_j| \leq C$$



Source: https://allmodelsarewrong.github.io/lasso.html

Setting $\widehat{\beta}_j = 0$: Pros and Cons

Pros:

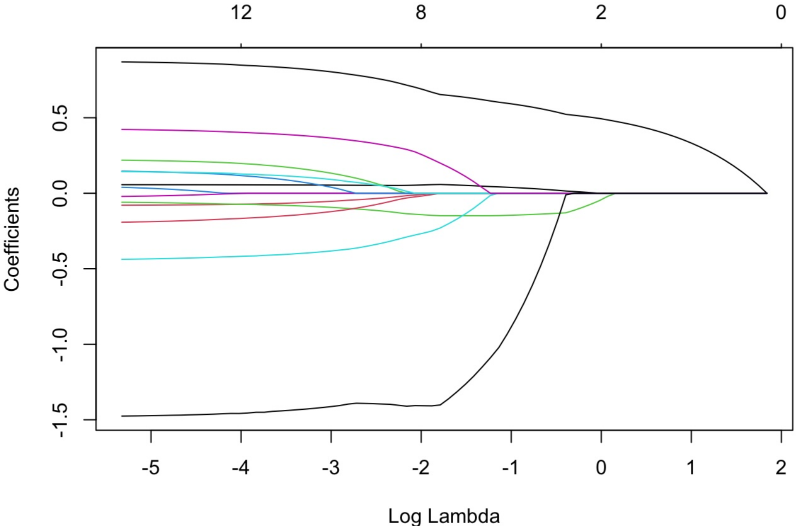
- Can interpret which variables are "most important"
- Data-driven way to exclude small coefficients to improve MSE and power
- Possible to run even if $p > n$

Cons:

- Always possible that picked model is "wrong"
- Estimator is biased
- Number of picked X depends on choice of λ

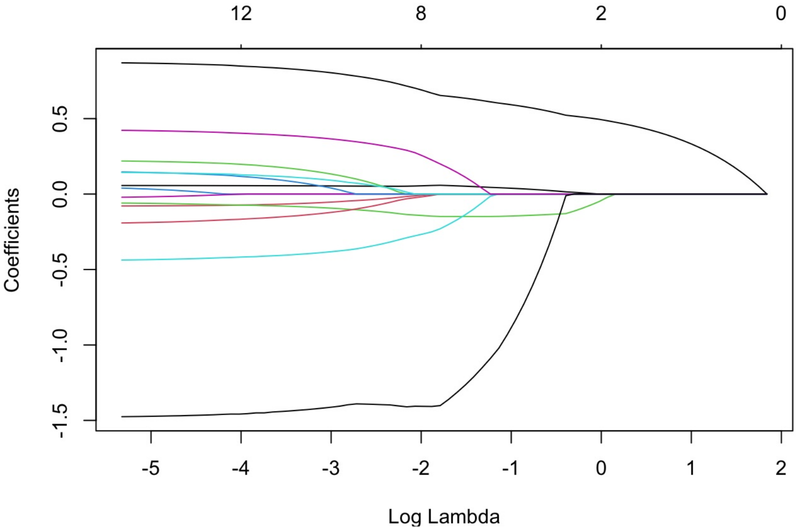
How Does λ Affect Estimator?

$\widehat{\beta} = \arg\min_b \sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \sum_{j=1}^{p} |b_j|$ ($\lambda \approx 0$: OLS. $\lambda \approx \infty$: all $\widehat{\beta}_j = 0$)



Source: https://bookdown.org/tpinto_home/Regularisation/lasso-regression.html

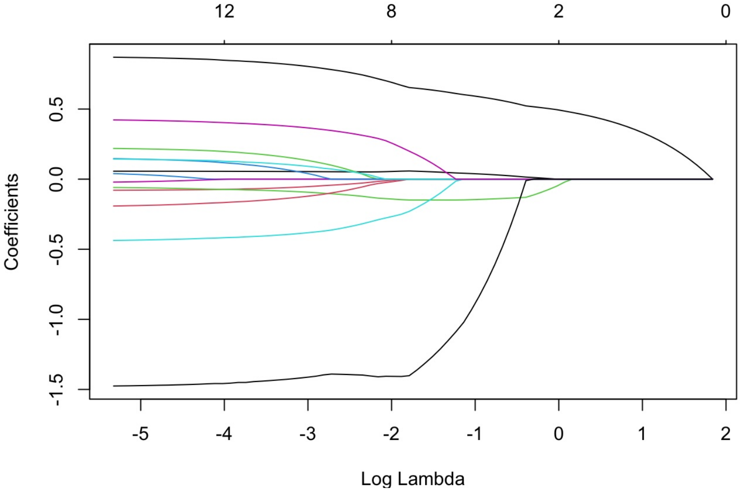
How Does λ Affect Estimator?



Source: https://bookdown.org/tpinto_home/Regularisation/lasso-regression.html

$\longrightarrow$ What are the "most important" variables?

Is There an Optimal λ?



- Extreme left: zero bias. Extreme right: zero variance.
- Some λ in the middle minimizes MSE!

Machine Learning and the "Truth"

- Remember $MSE = Bias^2 + Var$
  $\hookrightarrow$ Can pick λ to minimize MSE!

- But will that λ recover the "truth"/ "the true model"? Two points of view...

- Computer Scientist: "Truth? What is truth? Never heard of it... best λ gives the best model!"

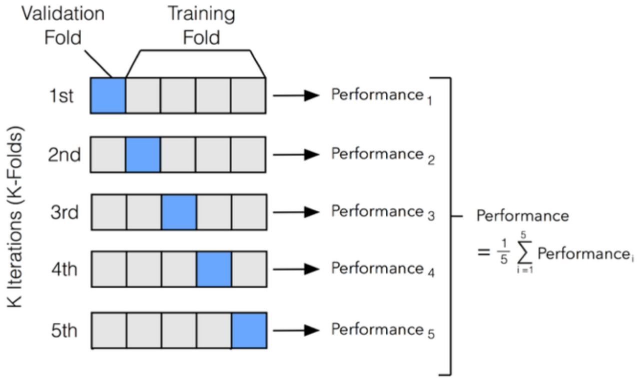
- Econometrician: "In general, no. But under conditions, yes!"

- Important to keep in mind what the goal is! Lasso as a tool for prediction vs Lasso as tool inference

- We will discuss a little bit of theory later

How To Select λ?

- Choosing λ is crucial for result of Lasso
- Data-driven way to pick λ: cross-validation
- Remember goal: minimize out-of-sample prediction error (MSE)!

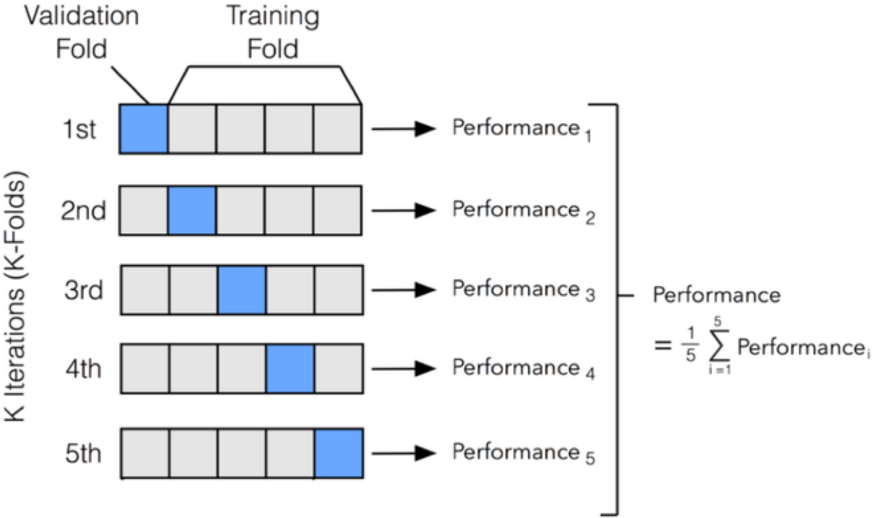


Source: https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-vaildation/

Cross Validation

1. Randomly split the sample of size n into K folds of size $\approx n_k$
2. Fix some value of λ
3. For $k = 1, \ldots, K$, learn $\widehat{\beta}_{-k}(\lambda)$ by running Lasso on data from all but fold k
4. For $k = 1, \ldots, K$, calculate the squared prediction error using fold k:
$$\Gamma_k(\lambda) = \sum_{i \in \mathcal{I}_k} (Y_i - X_i' \widehat{\beta}_{-k}(\lambda))^2$$
5. Aggregate total error as $\Gamma(\lambda) = \sum_{k=1}^{K} \Gamma_k(\lambda)$
6. Repeat steps 2 through 5, varying λ in a grid
7. Cross-validated λ given by $\widehat{\lambda} = \arg\min_{\lambda} \Gamma(\lambda)$

Cross Validation



Source: https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-vaildation/

Cross Validation

Cross validation has a major drawback

- Different choices of “seed” lead to different splits of the sample into folds ...
- ... and this will lead to different $\widehat{\beta}_{-k}(\lambda)$ and $\Gamma_k(\lambda)$
- Hence, picked λ will be different!

- Different choice of λ leads to different $\widehat{\beta}$ from final Lasso...
- ... and potentially even selected model will be different!!

- Suggestion (Wüthrich and Zhu, 2023): increase picked λ by 50%
  $\hookrightarrow$ In the context of Post-Double Lasso. In practice, can try other values too!

An Alternative Method for Choosing λ: BIC

- An alternative for Cross Validation is the BIC: Bayesian Information Criterion

$$\widehat{\lambda} = \arg\min_{\lambda} BIC(\lambda) := \log\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\widehat{\beta}(\lambda))^2\right) + |\widehat{S}(\lambda)| C_n \frac{\log(n)}{n}$$

where $\widehat{\beta}(\lambda)$ is Lasso estimator, and $|\widehat{S}(\lambda)|$ is number of X included by Lasso

- Advantage: no randomness from splitting the sample!

- Disadvantage: constant C_n is an arbitrary sequence "that tends to ∞"...
  $\hookrightarrow$ Additional hyperparameter to choose...

Is Lasso “Machine Learning”?

- Answer: depends on your definition of Machine Learning!

- Initial definition:
  1. Regularization for variance/bias tradeoff
     $\hookrightarrow$ In Lasso: λ!

  2. Data-driven procedure for choosing amount of regularization
     $\hookrightarrow$ In Lasso: cross-validation!
     - Is BIC “data-driven”?

A Little Bit of Theory...

Statistical Properties of the Lasso

- Important question: does the Lasso select the right model "asymptotically"?

- Yes, under the Irrepresentable Condition:

$$||\mathbb{E}[X_1X_1']^{-1}\mathbb{E}[X_1X_2']||_\infty \leq 1-\eta$$

  for some $\eta > 0$, where X_1 are the relevant variables, and X_2 the irrelevant
  $\hookrightarrow$ The coefficients of a regression of X_2 on X_1 are smaller than 1
  $\hookrightarrow$ That is, correlation between relevant and irrelevant not too high!

- Theorem (Zhao and Yu, 2006): Suppose p (n of covariates) and s (n of relevant covariates) are fixed, data is iid, X have finite second moments, and the irrepresentable condition holds. Furthermore, suppose that

$$\frac{\lambda_n}{n} \to 0 \text{ and } \frac{\lambda_n}{n^{\frac{1+c}{2}}} \to \infty \text{ for } 0 \leq c < 1$$

  then, the Lasso is model-selection consistent

Statistical Properties of the Lasso

- Even as n goes to infinity, Lasso only selects correct model under conditions
  - $\rightarrow$ Irrelevant variables not too correlated with relevant
  - $\rightarrow$ Asymptotics require specific rates for λ_n
- For finite n, performance can be worse
- Note: asymptotics assume number of coefficients is constant
- A lot of theory been done for p growing with n
  - $\hookrightarrow$ Number of covariates large **relative** to sample size
  - $\hookrightarrow$ Allows for $p > n$

Other Topics on Lasso

The Adaptive Lasso

1. Estimate β with ordinary Lasso

$$\widehat{\beta} = \arg\min_{b} \left(\sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda_1 \sum_{j=1}^{p} |b_j| \right)$$

   and let $\widehat{S}_1$ be set of selected covariates.

2. Estimate β by

$$\widetilde{\beta} = \arg\min_{b} \left(\sum_{i=1}^{n} \left(Y_i - \sum_{j \in \widehat{S}_1} X_{i,j} b_j \right)^2 + \lambda_2 \sum_{j \in \widehat{S}_1} |\widehat{\beta}_j|^{-1} |b_j| \right)$$

$\rightarrow$ Note: requires two tuning parameters λ_1 and λ_2

- $\hookrightarrow$ Can choose by cross-validation

The Adaptive Lasso

- Benefit: Adaptive Lasso is model-selection consistent (and oracle efficient)
- Compromise: requires $\lambda_{1,n}/\sqrt{n} \to \lambda^* > 0$, $\lambda_{2,n}/\sqrt{n} \to 0$, $\lambda_{2,n} \to \infty$

- Intuition: Lasso selects a model that is "too big"
- Adaptive Lasso penalizes more covariates with small $|\widehat{\beta}_j|$

- From theory perspective: Adaptive Lasso is "better"

Other Related Estimators

- There are other estimators that behave similarly to Lasso (model selection)
  - SCAD (smoothly clipped absolute deviation)
  - Bridge
  - Square-root Lasso
  - Elastic Net
- And penalized estimators that do not perform model selection, such as Ridge:

$$\widehat{\beta} = \arg\min_b \left(\sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \sum_{j=1}^{p} (b_j)^2 \right)$$

Post-Lasso / Post-Adaptive-Lasso

- Remember Lasso creates bias on $\widehat{\beta}$

$$\widehat{\beta} = \arg\min_{b} \left(\sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \sum_{j=1}^{p} |b_j| \right)$$

$\hookrightarrow$ There is shrinkage of $\widehat{\beta}$ towards zero because of the penalty!

- Alternative to bypass bias:
  1. Run Lasso/Adaptive-Lasso only to select covariates
  2. Run OLS (unbiased) on set of selected variables
- If model picked by Lasso is correct $\implies$ no bias in final estimator
- Post-Double-Lasso does something similar! (More on next talk)

Discussion

Discussion

- When would we want to run Lasso?
- How do we interpret output of Lasso?
- Is model selected by Lasso "optimal"?
- Does the Lasso select the "right" model?

Application Example

Example from Dizon-Ross and Jayachandran (2022)

- “This paper tests whether mothers and fathers differ in their spending on their daughters relative to their sons”

- Collect parents’ willingness to pay (WTP) for specific goods for their children
  - Becker-DeGroot-Marschak mechanism: “ask the respondent if he or she was willing to purchase the good at a series of prices, in declining order from the market price to a price near zero”

  - Incentivized (expensive): “after price questions, one price would be randomly chosen and she would purchase the good at that price if and only if her response had been that she wanted to”

  - Non-incentivized (cheaper): good was shown to respondent (concrete), but they knew no transaction would take place

- Can we pool incentivized and non-incentivized WTP?

Example from Dizon-Ross and Jayachandran (2022)

- "Appendix B presents evidence that the non-incentivized WTP elicitation appears to have worked quite similarly to the incentivized WTP elicitation. As a result, we pool incentivized and non-incentivized WTP in our main specifications for statistical power."
- Focus on one good (practice tests), which they asked in incentivized manner (I-WTP) (first round) and non-incentivized (N-WTP) (second round)

1. Split each sample (first and second rounds) in two, randomly
2. In first halves, run Lasso to select important predictors of I-WTP and N-WTP
   ↪ Same set of predictors! (2 among 29)
3. In second halves, run OLS on selected predictors and compare if coefficients are statistically different
   ↪ Differences are statistically insignificant!

Discussion

- Argument of the paper is that incentivized/non-incentivized WTP are similar
- Why split the sample?
- Why OLS in second stage?
- Under which conditions their approach is valid?
- Suggestions for improvement / robustness checks?
- Other thoughts?

Thank you!