Introduction to Machine Learning for Economists and Business Analysts

Anthony Strittmatter

Zoom

Zoom Link:

https://uni-wuerzburg.zoom.us/j/95131181830?pwd=ZTE3WEZXRW5pb3RBOERYd3VBckJjZz09

ID: 951 3118 1830

Password: 570475

Lecturer

Anthony Strittmatter

Research Interests: Business, Labour, and Health Economics, Program Evaluation, Computational Data Analytics

Positions:

Since 2020	Assistant Professor at the Institut Polytechnique in Paris, Center for Research in Economics and Statistics (CREST)
2014-2020	University of St.Gallen, with research visits at UC Berkeley, Stanford University, and Ludwig Maximilian University of Munich
2009-2016	Albert-Ludwig University of Freiburg

- Email: anthony.strittmatter@ensae.fr
- Webpage: www.anthonystrittmatter.com

Course Outline

- **Lecture 1:** Introduction to Statistical Learning
- **Lecture 2-4:** Supervised Machine Learning
  - Penalized Regression
  - Trees and Random Forest
- **Lecture 5:** Unsupervised Machine Learning
  - Clustering
  - Principal Component Analysis
- **Lecture 6:** Reinforcement Learning
  - Bandit Algorithm

Schedule

	Monday July 19	Tuesday July 20	Wednesday July 21	Thursday July 22	Friday July 23
3:15-4:45pm CEST	Lecture 1	Lecture 3	Lecture 4	Lecture 5	Lecture 6
5:15-6:45pm CEST	Lecture 2	PC Lab 1	PC Lab 2	PC Lab 3	PC Lab 4

PC Labs

- PC labs are integral part of the course.
- I will provide the entire course material on my GitHub repository: https://github.com/AStrittmatter/Wuerzburg
- I will use interactive Jupyter notebooks during the PC labs: https://mybinder.org/v2/gh/AStrittmatter/Wuerzburg/HEAD
- The advantage of the notebooks is that you do not need to install anything and the data is in the correct folder.
- However, in case the connection to the server is weak I also provide an R-file for download on my GitHub repository. To use it, you need to install R and RStudio on your PC. This are both open source softwares.

Grading

- Grading is based on an *individual* home assignment.
- Deadline for submission: Sunday, August 1, 2021

- It is allowed (and even desired) to support each other.
- But do not submit identical or copied home assignments, because then you will fail.

General

- Feel free to interrupt me at any time when you have questions.
- Tell me when I'm too slow or too fast. Ask me to repeat material in case something was not clear.
- You can also send me an email with questions: anthony.strittmatter@ensae.fr
- Proposals to improve the course are also welcome.
- Please try to interact as much as possible with your fellow students. Build learning groups.

References

- Mullainathan and Spiess (2017): "Machine Learning: An Applied Econometric Approach", Journal of Economic Perspectives, 31 (2), pp. 87-106, download.
- Athey (2017): "Beyond Prediction: Using Big Data for Policy Problems", Science, 355 (6324), pp. 483-485, download.

What is Machine Learning (ML)?

- ML (or statistical learning) methods exist already since decades.
- Currently "Machine Learning" is a buzz word
- Probably most people think of ML as some computational intensive methods that make data-driven modelling decisions and/or can deal with large data amounts.
- However, relevant textbooks consider even OLS/Logit as a statistical learning tool.

Purpose of Machine Learning

- Consider the structural model

$$Y = f(X) + \epsilon = X\beta + \epsilon,$$

with $E[\epsilon] = 0$.

- Causal analysis has the purpose to estimate $\hat{\beta}$, with $plim(\hat{\beta}) = \beta$.
- **Machine learning** has the purpose to predict Y.
- There is a clear link between causal analysis and machine learning, because

$$\widehat{Y} = \hat{f}(X) = X\hat{\beta}$$

is a potential predictor for Y.

- Parameter consistency has not the highest priority when it comes to predictions.

Potential Advantages and Disadvantages of ML

- ML methods can be very powerful to predict Y, even when $\hat{\beta}$ is biased.
- ML methods can incorporate many (or even high-dimensional) covariates X in a convenient way.
- ML methods can model $\hat{f}(\cdot)$ in a very flexible and data-driven way.
- **Main disadvantage:** ML is a black-box approach and we loose the interpretability of $\hat{f}(\cdot)$ or $\hat{\beta}$.

Prediction vs. Causality

Causal vs. Predictive Questions

Predictive Questions:

- How will the oil price change tomorrow (forecasting)?
- How high is the current unemployment rate (nowcasting)?
- Which adolescents have a high probability of becoming addicted to drugs (policy prediction)?

Causal Questions:

- What is the effect of a tweet by president Donald Trump on oil prices?
- How does inflation affect the unemployment rate?
- Can prevention programs reduce the probability of drug addiction among high risk youths?

Assessing the Model Accuracy

Causal Analysis:

- True β is unobservable.
- Assess the model with asymptotic properties

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2).$$

- Finite sample biases are mostly neglected.

Assessing the Model Accuracy

Prediction:

- We observe Y for each unit (e.g. individual).
- We can assess the model accuracy directly in the sample of our analysis, for example, using the mean-squared-error (MSE)

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2.$$

- MSE accounts for finite sample biases.

Example: Prediction of Used Car Prices

- We have access to web-scraped data from the online advertisement platform *myLemons*.
- We want to predict asking prices of used cars based on observable characteristics.
- We observe around 40 covariates about car brand, mileage, age, emissions, maintenance certificate, seller type, guarantee, etc. (including several non-linear and interaction terms)

In-Sample MSE

- Partition data into training and test sample
- In the training sample, we estimate the empirical model

$$Y_{tr} = \hat{f}_{tr}(X_{tr}) + \hat{\epsilon}_{tr} = X_{tr}\hat{\beta}_{tr} + \hat{\epsilon}_{tr}$$

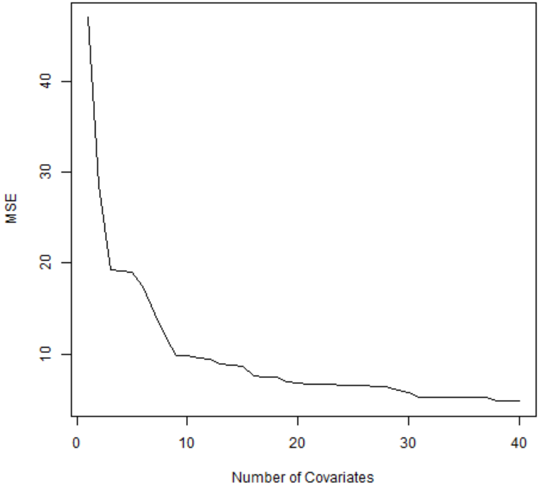
- In the training sample, we predict the fitted values

$$\widehat{Y}_{tr} = \hat{f}_{tr}(X_{tr}) = X_{tr}\hat{\beta}_{tr}$$

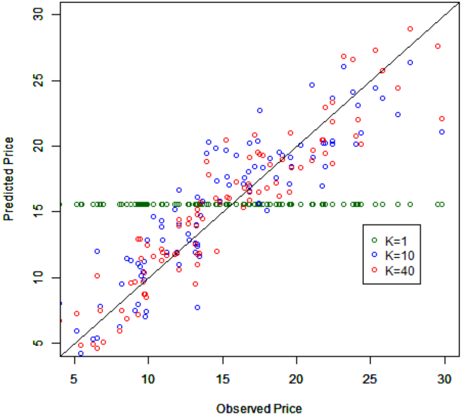
  and calculate the MSE

$$\widehat{MSE}_{tr} = \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} (Y_{i,tr} - \widehat{Y}_{i,tr})^2.$$

MSE in Training Sample



Predicted Car Prices in Training Sample



Number of Covariates	1	10	40
MSE	46.948	9.819	4.866

Out-of-Sample MSE

- In the training sample, we estimate the empirical model

$$Y_{tr} = \hat{f}_{tr}(X_{tr}) + \hat{\epsilon}_{tr} = X_{tr}\hat{\beta}_{tr} + \hat{\epsilon}_{tr}$$

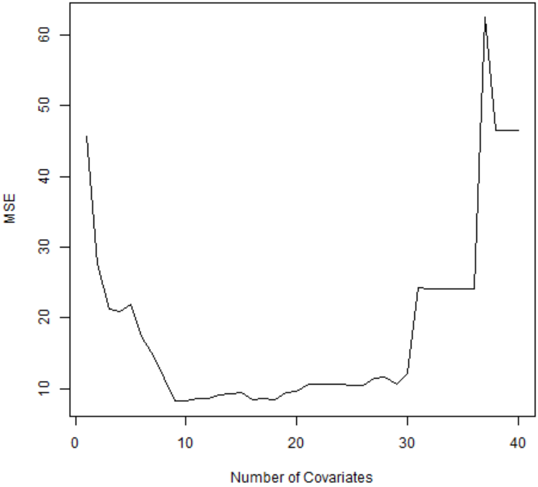
- In the test sample, we predict the fitted values

$$\widehat{Y}_{te} = \hat{f}_{tr}(X_{te}) = X_{te}\hat{\beta}_{tr}$$

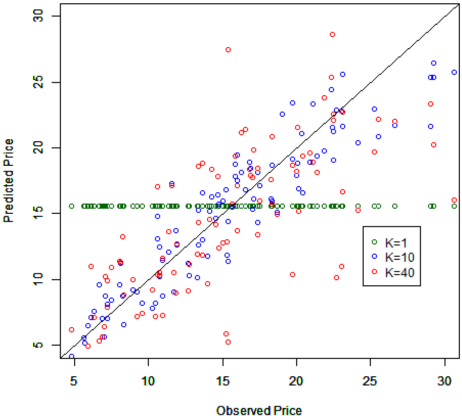
and calculate the MSE

$$\widehat{MSE}_{te} = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} (Y_{i,te} - \widehat{Y}_{i,te})^2.$$

MSE in Test Sample



Predicted Car Prices in Test Sample



Number of Covariates	1	10	40
MSE	45.742	8.222	46.499

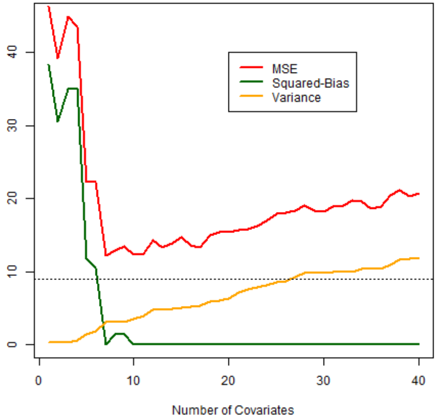
Bias-Variance Trade-Off

- When we assess the model for one randomly drawn individual from the test sample with fixed characteristics x_{te}, then we can decompose the MSE to

$$
\begin{aligned}
MSE_{te} =& E[(Y_{te} - \widehat{Y}_{te})^2] \\
=& E[(f(x_{te}) + \epsilon_{te} - \hat{f}_{tr}(x_{te}))^2] \\
=& \underbrace{E[(f(x_{te}) - \hat{f}_{tr}(x_{te}))^2]}_{\text{Reducible}} + \underbrace{Var(\epsilon_{te})}_{\text{Irreducible}} \\
=& \underbrace{E[f(x_{te}) - \hat{f}_{tr}(x_{te})]^2}_{\text{Squared-Bias}} + \underbrace{Var(\hat{f}_{tr}(x_{te}))}_{\text{Variance}} + Var(\epsilon_{te})
\end{aligned}
$$

- For i.i.d. data, $\hat{f}_{tr}(\cdot)$ and ϵ_{te} are independent of each other.

Simulation of Bias-Variance Trade-Off



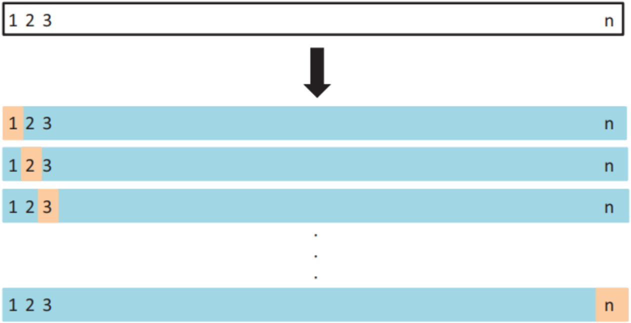
- Only the first ten covariate have an impact on car prices in the simulation.
- Horizontal dashed line is the simulated noise $Var(\epsilon_{te})$.

Lasso Example

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

	OLS	Lasso
Intercept	21.246	22.776
diesel	2.075	.
other_car_owner	0.730	.
pm_green	1.635	.
private_seller	6.100	0.076
guarantee	-2.440	-0.437
inspection	-0.813	.
maintenance_cert	1.481	.
mileage	-0.049	-0.031
age_car_years	-1.291	-1.012
R^2 training	0.655	0.543
R^2 test	0.606	0.611

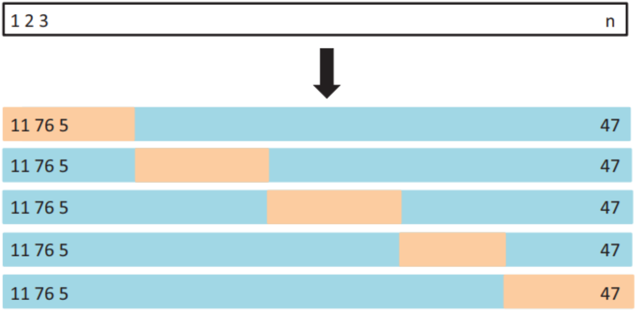
Leave-One-Out Cross-Validation



Source: James et al. (2013), p. 179

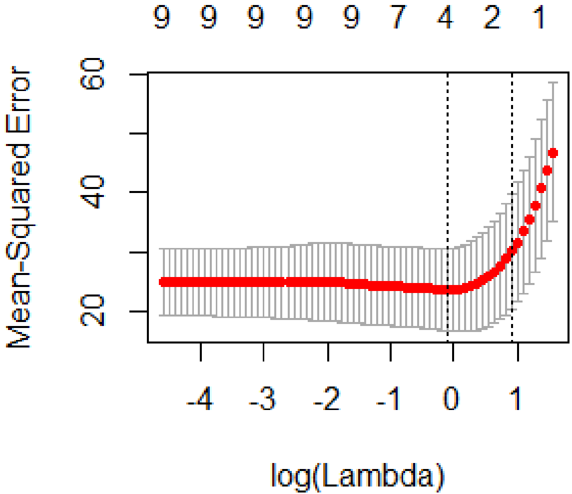
Selection of Optimal Penalty Parameter

k-fold Cross-Validation (CV) Algorithm



Source: James et al. (2013), p. 181

Cross-Validated MSE



Stability of the Lasso Model

	Lasso 1	Lasso 2	Lasso 3	Lasso 4	Lasso 5
Intercept	22.776	25.947	24.937	27.309	25.116
diesel	.	.	2.387	.	0.886
other_car_owner	.	-1.257	0.393	.	.
pm_green	.	2.871	.	.	.
private_seller	0.076	5.094	.	-1.037	.
guarantee	-0.437	1.677	15.939	.	.
inspection	.	-0.666	-0.374	.	.
maintenance_cert	.	-2.579	-0.868	.	.
mileage	-0.031	-0.037	-0.041	-0.069	-0.062
age_car_years	-1.012	-1.347	-1.416	-0.874	-1.115

→ We do not learn the "true" structural model from ML

→ ML is a black-box approach

Stability of the Lasso Predictions

Correlation of Predicted Car Prices in Test Sample:

	Lasso 1	Lasso 2	Lasso 3	Lasso 4
Lasso 2	0.94			
Lasso 3	0.85	0.81		
Lasso 4	0.97	0.91	0.85	
Lasso 5	0.99	0.94	0.87	0.99

When could Predictions be Useful?

Tasks with a prediction purpose:

- Predict stock or commodity prices using Twitter data.
- Nowcasting unemployment rate or GDP using Google search queries.
- Pre-screening of job applications.
- Consumer demand (shipping before the order occurs).
- Movie recommendations on Netflix.
- Handwriting, image, face, or voice recognition.
- Generating data

Examples of Business and Economic Studies

Prediction Tasks:

- Chandler, Levitt, and List (2011) predict shootings among high-risk youth to target mentoring interventions.
- Kleinberg, et al. (2018) predict the crime probability of defendants released from investigative custody to improve judge decisions.

Pre-Processing Unstructured Data:

- Glaeser et al. (2016) use images from Google Street View to measure block-level income in New York City and Boston.
- Kang et al. (2013) use restaurant reviews on Yelp.com to predict the outcome of hygiene inspections.
- Kogan et al. (2009) predict volatility of firms from market-risk disclosure texts (annual 10-K forms).