Advanced Empirical Finance: Topics and Data Science

Stefan Voigt

Spring 2024

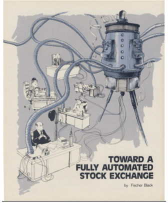
University of Copenhagen and Danish Finance Institute

Machine learning

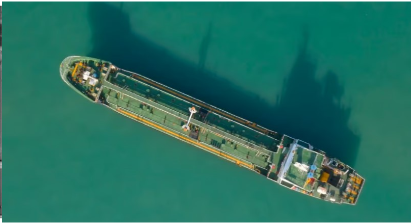
What is Machine learning?

*The definition of "machine learning" is inchoate and is often context specific. We use the term to describe **(i)** a diverse collection of high-dimensional models for statistical prediction, combined with **(ii)** so-called "regularization" methods for model selection and mitigation of overfit and **(iii)** efficient algorithms for searching among a vast number of potential model specifications. (Gu et al. 2020)*

- (i) select between small simplistic and complex ML models
- (i) Focus on predictive accuracy
- (ii) selecting from multiple models in-sample leads to overfitting and poor out-of-sample performance
- (ii) "regularization" methods for model selection
- (iii) challenge in terms of computational effort



ML in Finance



What makes ML in Finance special?

Challenges (Israel, Kelly, Moskowitz, 2019)

- Limited data (left-hand side limited by T)
- Markets evolve and thus even lower effective sample size
- By market efficiency: small signal-to-noise ratio (limited predictability)
- Data potentially unstructured (company announcements)

Our aim

- Exploit potential for improving risk premium measurement $E_t\left(r_{i,t+1}\right)$

But...

- improved predictions are still only measurements
- The measurements do not tell us about economic mechanisms or equilibria
- Machine learning methods on their own do not identify fundamental associations among asset prices and conditioning variables

Overview: Empirical Asset Pricing via Machine Learning

- Familiarize yourself with the paper "Empirical Asset Pricing via Machine Learning" by Gu et al. (2020)
- comparative analysis of machine learning methods for the canonical problem of measuring asset risk premiums
- "We demonstrate large economic gains to investors using *machine learning forecasts*, in some cases doubling the performance of leading regression-based strategies from the literature."

Machine learning roadmap

1. Bias-Variance Trade-off
2. Penalized Linear Regressions (Ridge and Lasso)
3. Regression Trees and Random Forests
4. Neural Networks
5. Advanced case studies and applications

Your task:

- Return prediction for all CRSP-listed stocks
- Large set of macroeconomic predictors
- Hundreds of predictive firm and economic characteristics
- You should study Gu et al. (2020) in depth!
- **Exercises:** Prepare the dataset as explained in Section 2.1 of Gu et al. (2020)

Bias-Variance Trade-off

Unbiased, linear estimators

$$E_t\left(r_{i,t+1}\right) = g(x_{i,t}) \stackrel{??}{=} \beta' x_{i,t}$$

- Machine learning prescribes a vast collection of high-dimensional models that attempt to predict future quantities of interest while imposing regularization
- We know: OLS is the best linear unbiased estimator (BLUE)
- "Best" = the lowest variance estimator among all other *unbiased linear* estimators
- Requiring the estimator to be *linear* is binding since *nonlinear* estimators exist (e.g., neural networks or regression trees)
- Likewise, *unbiased* is crucial since *biased* estimators do exist

Biased estimators?

- *Shrinkage* methods: the variance of the OLS estimator can be high as OLS coefficients are unregulated
- If judged by Mean Squared Error (MSE), biased estimators could be more attractive if they produce substantially smaller variance than OLS

Shortcomings of OLS

- Let β denote the true regression coefficient and let $\hat{\beta} = (X'X)^{-1} X'y$, where X is a $(T \times N)$ matrix of explanatory variables
- Then, the variance of the (unbiased) OLS estimate $\hat{\beta}$ is given by

$$\begin{aligned} Var\left(\hat{\beta}\right) &= E\left(\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right) \\ &= E\left((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right) \\ &= \sigma_\varepsilon^2 E\left((X'X)^{-1}\right) \end{aligned}$$

  where ε is the vector of residuals and σ_ε^2 is the variance of the error term
- When the predictors are highly correlated, the term $(X'X)^{-1}$ quickly explodes
- Even worse: the OLS solution is not unique if X is not of full rank

OLS in a prediction context

1. restrictive
2. may provide poor predictions, may be subject to *over-fitting*
3. does not penalize for model complexity and could be difficult to interpret

The Bias-Variance Trade-off

- Assume the model

$$y = f(x) + \varepsilon, \quad \varepsilon \sim (0, \sigma_\varepsilon^2)$$

- β^{ols} has a host of well-known properties (Gauss-Markov)
- But: Can we choose $\hat{f}(x)$ to fit future observations well?
- MSE depends on the model as follows:

$$\begin{aligned}
E(\hat{\varepsilon}^2) &= E((y - \hat{f}(\mathbf{x}))^2) = E((f(\mathbf{x}) + \varepsilon - \hat{f}(\mathbf{x}))^2) \\
&= \underbrace{E((f(\mathbf{x}) - \hat{f}(\mathbf{x}))^2)}_{\text{total quadratic error}} + \underbrace{E(\varepsilon^2)}_{\text{irreducible error}} \\
&= E\left(\hat{f}(\mathbf{x})^2\right) + E\left(f(\mathbf{x})^2\right) - 2E\left(f(\mathbf{x})\hat{f}(\mathbf{x})\right) + \sigma_\varepsilon^2 \\
&= E\left(\hat{f}(\mathbf{x})^2\right) + f(\mathbf{x})^2 - 2f(\mathbf{x})E\left(\hat{f}(\mathbf{x})\right) + \sigma_\varepsilon^2 \\
&= \underbrace{\mathrm{Var}\left(\hat{f}(\mathbf{x})\right)}_{\text{variance of model}} + \underbrace{E\left((f(\mathbf{x}) - \hat{f}(\mathbf{x}))\right)^2}_{\text{squared bias}} + \sigma_\varepsilon^2
\end{aligned}$$

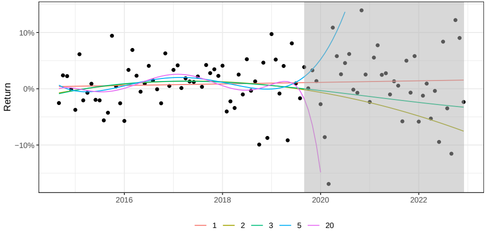
- A biased estimator with small variance may have a lower MSE than an unbiased estimator

Over-fitting example

- 100 monthly manufacturing industry excess returns
- Estimate a polynomial regression

$$r_t = \alpha + \sum_{p=1}^{P} \beta_p t^p$$

  where t is a time index, ranging from 1 to 60
- Evaluate the performance in-sample and out-of-sample for $P = 1, 2, 3, 5, 20$



Ridge Regression

- Introduced by Hoerl and Kennard (1970a, 1970b)
- Impose a penalty on the L_2 norm of the parameters $\hat{\beta}$ such that for $c \geq 0$ the estimation takes the form

$$\beta^{\text{ridge}} = \arg\min_{\beta} (y - X\beta)' (y - X\beta) \;\; \text{s.t.} \;\; \beta'\beta \leq c$$

- Standard optimization procedure yields

$$\beta^{\text{ridge}} = (X'X + \lambda I)^{-1} X'y$$

- Hyper parameter λ (c) controls the amount of regularization
- Note that $\beta^{\text{ridge}} = \beta^{\text{ols}}$ for $\lambda = 0$ ($c \to \infty$) and $\beta^{\text{ridge}} \to 0$ for $\lambda \to \infty$ ($c \to 0$)
- ($X'X + \lambda I$) is non-singular even if $X'X$ is
- *Note:* Usually, the intercept is not penalized (in practice: demean y)

Ridge Regression

- Let $D := X'X$

$$\begin{aligned}\beta^{\text{ridge}} &= (X'X + \lambda I)^{-1} X'y \\ &= (D + \lambda I)^{-1} DD^{-1}X'y \\ &= \left(D\left(I + \lambda D^{-1}\right)\right)^{-1} D\beta^{\text{ols}} \\ &= \left(I + \lambda D^{-1}\right)^{-1} D^{-1}D\beta^{\text{ols}} = (I + \lambda D)^{-1} \beta^{\text{ols}}\end{aligned}$$

- β^{ridge} is biased because $E(\beta^{\text{ridge}} - \beta) \neq 0$ for $\lambda \neq 0$
- *But* at the same time (under homoscedastic error terms)

$$\text{Var}(\beta^{\text{ridge}}) = \sigma_\varepsilon^2 (D + \lambda I)^{-1} X'X (D + \lambda I)^{-1}$$

- You can show that $\text{Var}(\beta^{\text{ridge}}) \leq \text{Var}(\beta^{\text{ols}})$
- Trade-off between bias and variance of the estimator!