# Statistical Learning

Academic year 2022/23
CLAMSES - University of Milano-Bicocca

Aldo Solari
aldo.solari@unimib.it

# Webpages

MOODLE: https://elearning.unimib.it/course/view.php?id=44902

- Syllabus
- Forum
- Grades
- Exercises

WEB: https://aldosolari.github.io/SL/

- Calendar
- Slides, R code, exercises
- Textbooks
- Exam

# Exam

The exam consists in a written examination (and an optional oral examination).

The written (open-book) examination will be held in lab.

- Questions about theory
- Computational exercises
- Data analysis tasks

# Program

- Prediction
  - Conformal prediction
- Estimation
  - James-Stein estimation
  - Ridge regression
  - Smoothing splines
  - Sparse modeling and the Lasso
- Attribution
  - Data splitting for variable selection
  - Stability Selection
  - Knockoff filter
  - Leave-one-covariate-out (LOCO) inference

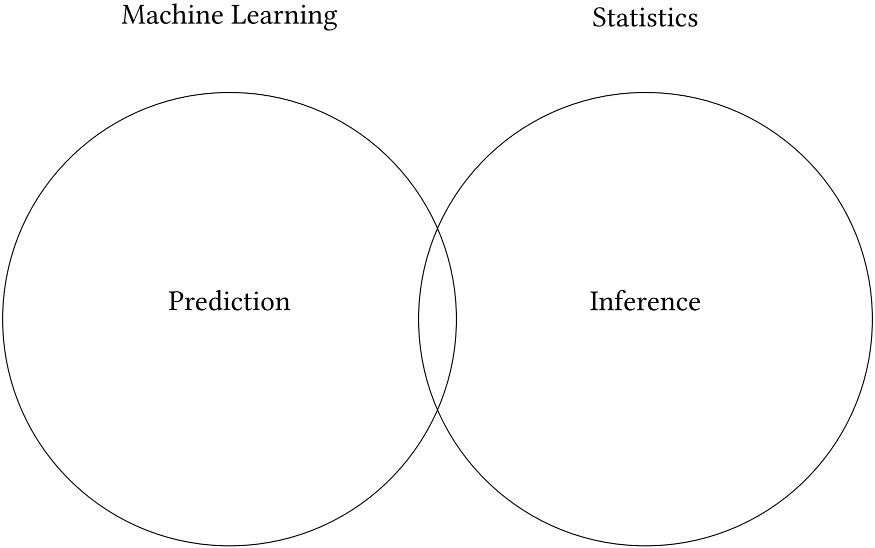
# Table of Contents

I. Prediction

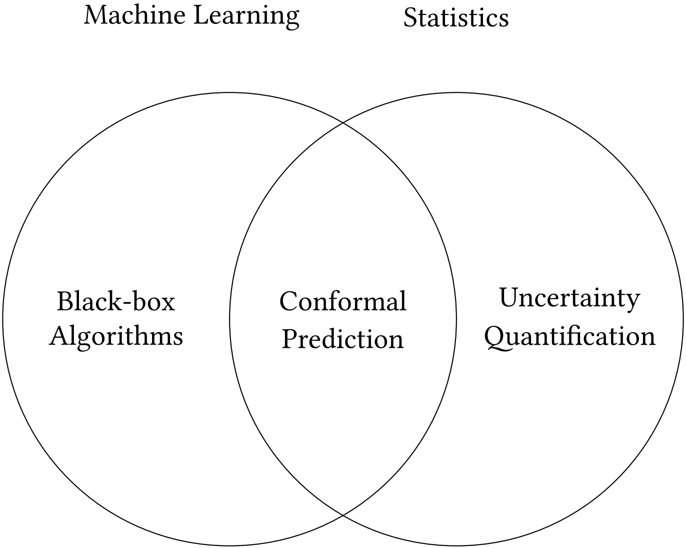
II. Estimation

III. Attribution

2001

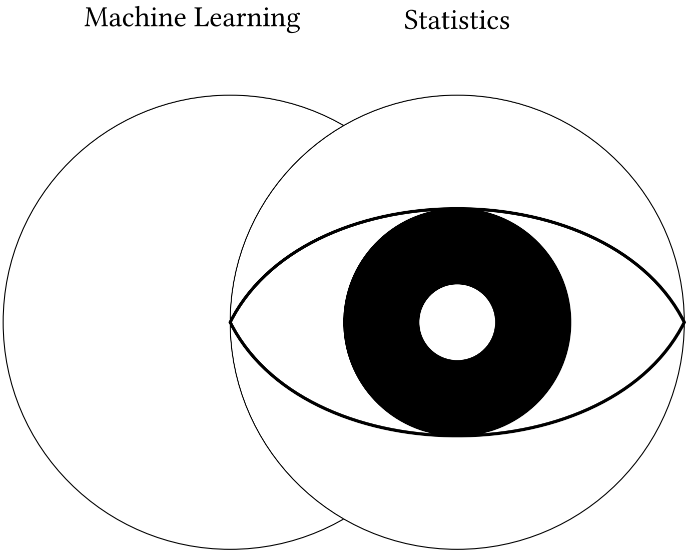


2023



Machine Learning
Statistics
Black-box Algorithms
Conformal Prediction
Uncertainty Quantification

# Statistical Point of View



Modern prediction algorithms such as random forests and deep learning use training sets, often very large ones, to produce rules for predicting new responses from a set of available predictors.

A second question—right after "How should the prediction rule be constructed?"—is "How accurate are the rule's predictions?"

From: Efron, B. (2021). Resampling plans and the estimation of prediction error. Stats, 4(4), 1091-1115.

How to quantify the uncertainty of predictions from algorithms used in machine learning ?

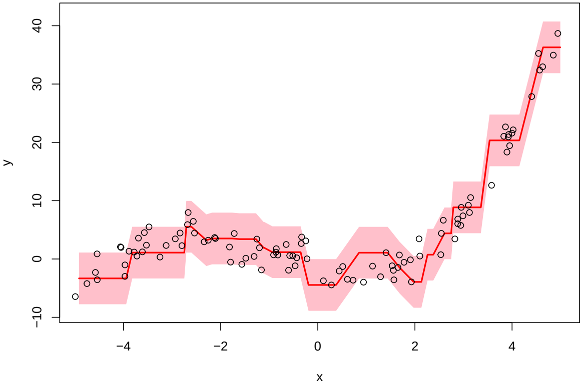
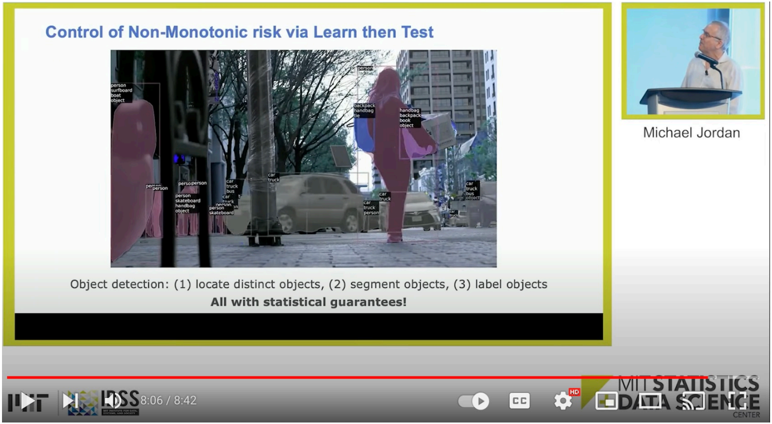



***Figure 1: Prediction set examples on Imagenet.*** *We show three progressively more difficult examples of the class* **`fox squirrel`** *and the prediction sets (i.e., $\mathcal{C}(X_{\text{test}})$) generated by conformal prediction.*

From: Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.




From: Michael I. Jordan on Conformal Prediction

https://www.youtube.com/watch?v=kSGP4F_ZcBY

# Table of Contents

I. Prediction

II. Estimation

III. Attribution

# James-Stein estimation

Let $X_1$, $X_2$ and $X_3$ be independent r.v. with $X_i \sim N(\mu_i, 1)$.

Writing $X = (X_1, X_2, X_3)$, suppose we want to find a good estimator $\hat{\mu} = \hat{\mu}(X)$ of $\mu = (\mu_1, \mu_2, \mu_3)$

Squared error loss function:

$$L(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|^2 = (\hat{\mu}_1 - \mu_1)^2 + (\hat{\mu}_2 - \mu_2)^2 + (\hat{\mu}_3 - \mu_3)^2$$

Risk function: $R(\hat{\mu}, \mu) = \mathbb{E}[L(\hat{\mu}, \mu)]$

MLE is $\hat{\mu} = X$. $\exists$ an alternative estimator $\tilde{\mu}$ such that $R(\tilde{\mu}, \mu) \leq R(\hat{\mu}, \mu)$ for all $\mu$, with strict inequality for some value of $\mu$?

# Ridge regression

- The ML estimator of the parameter of the linear regression model $\hat{\beta} = (X^t X)^{-1} X^t y$ is only well-defined if $(X^t X)^{-1}$ exists.
- In wide-data situations where $p > n$, the rank of $X^t X$ is $n < p$, and, consequently, it is singular. Hence, the regression parameter $\beta$ cannot be estimated.
- How to perform high-dimensional regression?

# Smoothing splines

`mcycle` dataset (`MASS` R package), gives $n = 133$ observations of accelerometer readings taken through time (after impact) in an experiment on the efficacy of crash helm

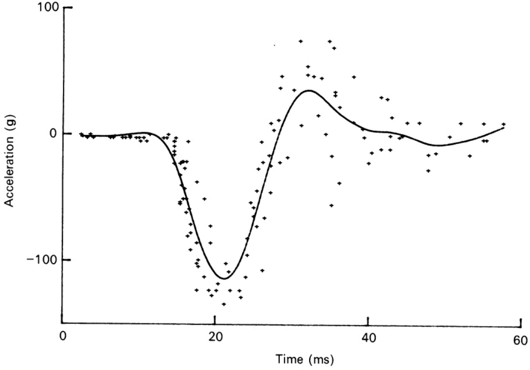



Fig. 3. The motor-cycle impact data with automatically chosen smoothing curve.

From: Silverman (1985) Some aspects of the spline smoothing approach to non-parametric curve fitting. JRSS-B, 47:1-52.

# Classical vs high-dimensional theory

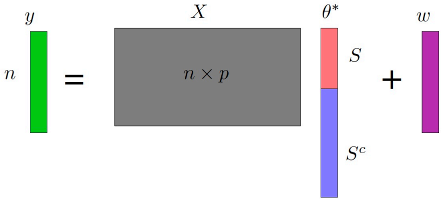
- Consider Linear Discriminant Analysis where the two classes are distributed as $p$-variate Gaussians $X_1 \sim N(\mu_1, I_p)$ and $X_2 \sim N(\mu_2, I_p)$ with $\gamma = \|\mu_1 - \mu_2\|$
- Classical theory: if $(n_1, n_2) \to \infty$ and $p$ remains fixed, then LDA error probability $\stackrel{prob.}{\to} \Phi(-\gamma/2)$
- High-dimensional theory: if $(n_1, n_2, p) \to \infty$ with $p/n_i \to \delta$, then LDA error probability $\stackrel{prob.}{\to} \Phi\left(-\frac{\gamma^2}{2\sqrt{\gamma^2+2\delta}}\right)$
- LDA error probability for

$$(p, n_1, n_2) = (400, 800, 800)$$

  is better described by the classical or the high-dimensional theory? e.g. for $\gamma = 1$ and $\delta = 1/2$, LDA error probability $\approx 31\%$ (classical) or $\approx 36\%$ (high-dimensional)?

# Sparse modeling and the Lasso

A sparse statistical model is one having only a small number of nonzero parameters (easier to estimate and interpret)




**Set-up:** noisy observations $y = X\theta^* + w$ with sparse $\theta^*$

Source: M.J. Wainwright

The best subset selection (variable selection) problem is nonconvex and NP-hard. The lasso (Tibshirani, 1996) [cited by 51K] solves a convex relaxation of it by replacing the $\ell_0$ norm by the $\ell_1$ norm.

# Table of Contents


I. Prediction

II. Estimation

III. Attribution

# Data splitting

```
library(tidyverse)
library(ISLR)
dataset <- Hitters %>% na.exclude
n <- nrow(dataset)
set.seed(123)
dataset$Salary <- rexp(n, 1/mean(dataset$Salary))
summary(stepAIC(lm(Salary ~ ., dataset), trace=F))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 466.65825  102.36325   4.559 7.96e-06 ***
AtBat         0.51870    0.33543   1.546   0.1232
Walks        -4.50902    2.54583  -1.771   0.0777 .
CAtBat       -0.08607    0.04093  -2.103   0.0364 *
CWalks        0.82056    0.38464   2.133   0.0338 *
LeagueN     149.31154   63.22722   2.362   0.0189 *
```

# Stability selection

Not a new variable selection technique, it improves existing methods

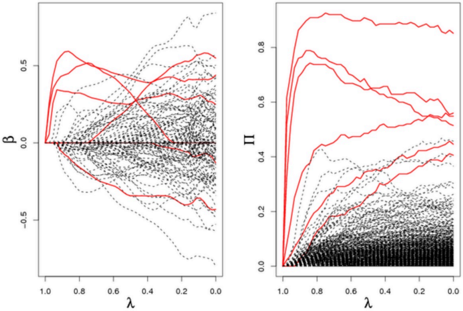
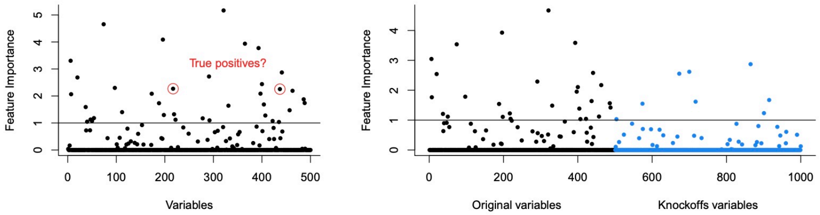



Figure 1 from Meinshausen and Bühlmann (2010)
regularisation and stability path for the lasso

# Knockoff filter

How to control the false discovery rate when performing variable selection?




Source: E. Candés

# Textbooks

- Efron, Hastie (2016) Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press [CASI]
- Hastie, Tibshirani, Friedman (2009). The Elements of Statistical Learning. Springer [ESL]
- Hastie, Tibshirani, Wainwright (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press [SLS]
- Lewis, Kane, Arnold (2019) A Computational Approach to Statistical Learning. Chapman And Hall/Crc. [CASL]
- Wainwright (2019) High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press [HDS]