Statistical Learning

Academic year 2021/22
CLAMSES - University of Milano-Bicocca

Aldo Solari
aldo.solari@unimib.it

Webpages

MOODLE: https://elearning.unimib.it/course/view.php?id=38049

- Syllabus
- Forum
- Grades

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

In Data Mining we have discussed Prediction.

- Estimation
  - James-Stein estimation
  - Ridge regression
  - Smoothing splines
  - Classical versus high-dimensional theory
  - Sparse modeling and the Lasso
  - Best Subsets Selection
- Attribution
  - Data splitting for variable selection
  - Stability Selection
  - Knockoff filter
  - Conformal prediction

James-Stein estimation

Suppose that we were interested in estimating

- μ_1: the US wheat yield for 1993
- μ_2: the number of spectators at the Wimbledon tennis tournament in 2001
- μ_3: the weight of a randomly chosen candy bar from the supermarket.

Suppose we have independent Gaussian measurements $X_1 \sim N(\mu_1, 1)$, $X_2 \sim N(\mu_2, 1)$ and $X_3 \sim N(\mu_3, 1)$ of each of these quantities.

Does make sense that the estimate of the US wheat yield depends on the number of spectators at Wimbledon and the weight of a candy bar? i.e. $\hat{\mu}_1 = \hat{\mu}_1(X_1, X_2, X_3)$?

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 β 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

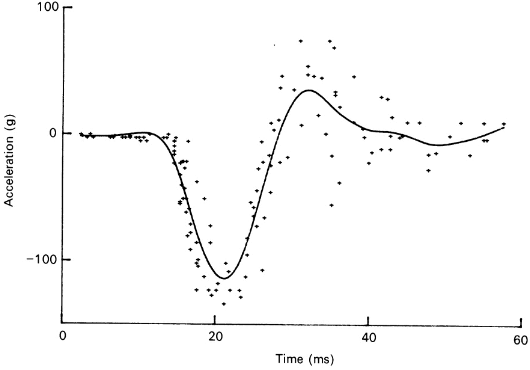


Fig. 3. The motor-cycle impact data with automatically chosen smoothing curve.

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

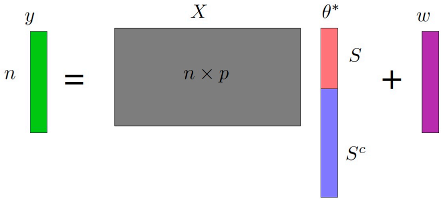
- Consider Linear Discriminant Analysis where the two classes are distributed as p-variate Gaussians $X_1 \sim N(\mu_1, I_p)$ and $X_2 \sim N(\mu_2, I_p)$ with $\gamma = \|\mu_1 - \mu_2\|$
- Classical theory: if $(n_1, n_2) \to \infty$ and p remains fixed, then LDA error probability $\overset{prob.}{\to}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

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 θ^*

Source: M.J. Wainwright

The best subset selection (variable selection) problem is nonconvex and NP-hard. The lasso (Tibshirani, 1996) [cited by 48K] solves a convex relaxation of it by replacing the ℓ_0 norm by the ℓ_1 norm.

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

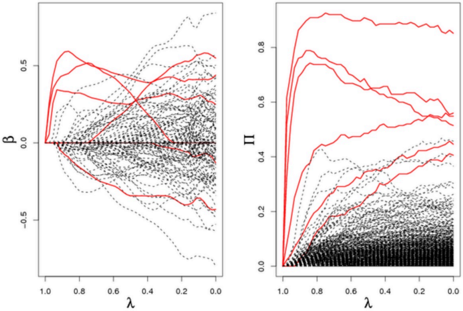
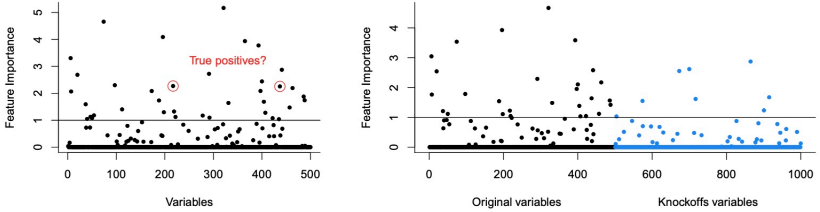


Figure 1 from Meinshausen and Bühlmann (2010)
regularisation and stability path for the lasso

Knockoff filter

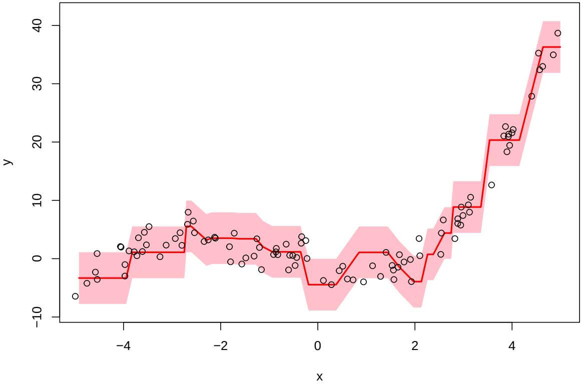
How to control the false discovery rate when performing variable selection?



Source: E. Candés

Conformal prediction

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



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]
- Wasserman (2006) All of Nonparametric Statistics. Springer [ANS]