Machine Learning for All: Examples for Subset Selection & Lasso

Anastasiya Yarygina

Monday, January 21, 2019

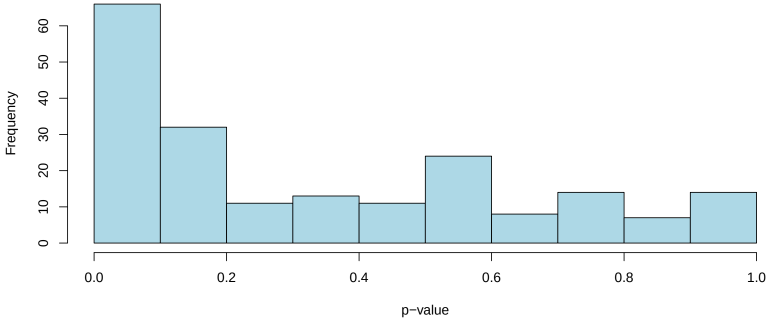
Practical Excercises

- Subset selection
  - False Discovery Rate (FDR)[1] as a selection tool
  - In-sample and Out of Sample (OOS) fit
  - Forward stepwise Regression
- Regularization
  - LASSO Regularization Path
  - Parameter selection using Cross Validation
- Data: Semiconductors dataset

[1] FDR is the expected proportion of false positives

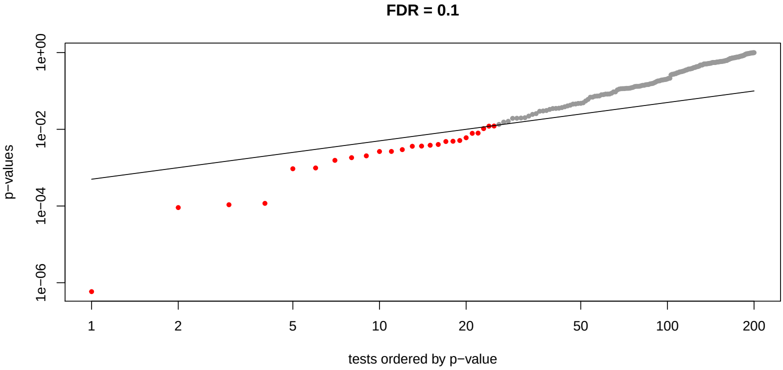
Semiconductors dataset: explore predictors

- This dataset has 201 predictors
- Some p-values are clustered at zero. But which are **significant signals**?



Select predictors using FDR

- How many predictors are in fact good signals (q=10% FDR)?




```
## The nubmer of significant signals:

## [1] 25
```

Compare **in-sample** fit of **full** and **cut** models

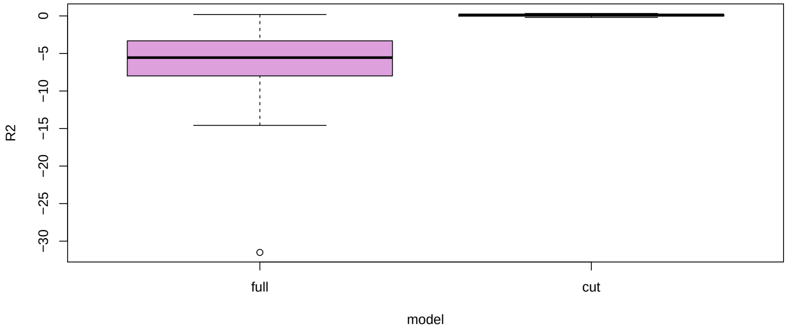
- Fit a new **cut** model using only 25 best signals
- How does the in-sample fit change?

```
## Full model R2= 0.5621432
```

```
## Cut model R2= 0.1811822
```

Compare **OOS** fit of **full** and **cut** models

Split data in 10 random samples, fit **full** and **cut** models on 9 samples, predict on 10th. What are the average R^2?



Cut model mean OOS R^2 is about $1/2$ in-sample R^2. Full model is terrible!

Forward Stepwise Regression

1. Fit all univariate models. Choose the one with the highest R^2, keep this variable, say X1, for your final model.
2. Fit all bivariate models including X1, choose the one with the highest R^2, keep two variables, say, X1 and X15.
3. Repeat: max R^2 by adding one variable at time to the model.
4. Stop when AIC is lower for the current model than for any of the models that add one variable.

The Forward Stepwise procedure chooses around 70 coefficients.

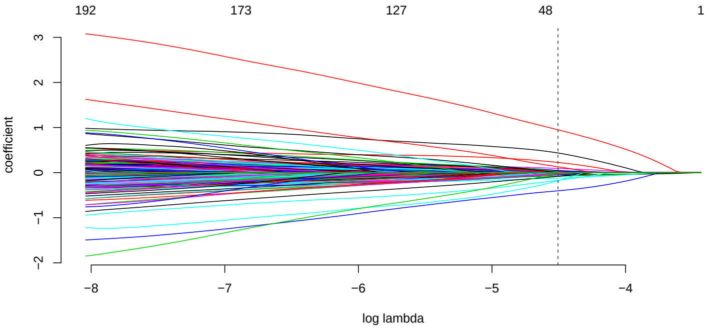
Regularization using LASSO[2]

- Depart from optimality:
  - minimize deviance + **cost on an absolute size** of coefficients
- By penalizing we **shrink** some **estimates towards zero**
- Some coefficients can become zero and get eliminated from the model
- Tunning parameter λ is the **amount of shrinkage**

[2] Least Absolute Shrinkage and Selection Operator

LASSO Algorithm using *gamlr*[3] package

- Start with large λ_1 so that $\hat{\beta} = 0$
- For $t = 2 \ldots T$ update $\hat{\beta}$ to be optimal under $\lambda_t < \lambda_{t-1}$



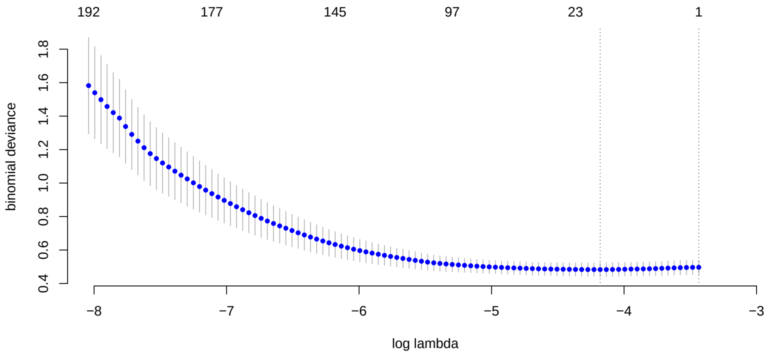
At the top of the figure: number of non-zero coefficients

[3] Gamma-Lasso regression

Cross Validation using *gamlr*

- Set a sequence of $\lambda_1, ..., \lambda_T$
- For each $k = 1, ..., K$ folds:
  - Fit the path on all data except fold k
  - Get fitted deviance on left-out data
- Select λ that gives minimum average OOS deviance

Cross Validation using *gamlr*



Compare AICc selection and Cross Validation selection

- Compare log(*lambda*) under different selection criteria

```
## AICc:  -4.50608

## AICc chooses  31 coefficients

## ----------------------------------------

## Min deviance:  -4.180462

## Min deviance chooses 12 coefficients
```