Recap 7

Regularization

ISLR 6, ESL 3

Jilles Vreeken
Krikamol Muandet

UNIVERSITÄT DES SAARLANDES

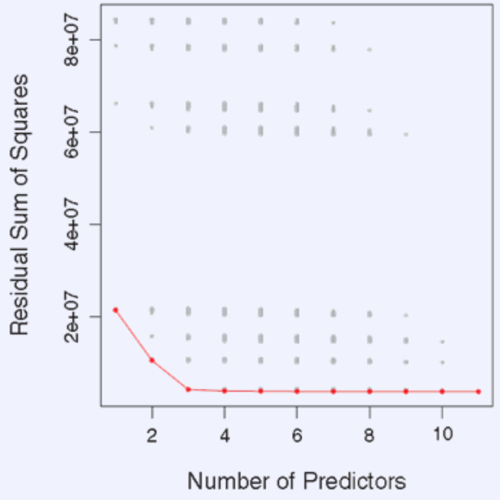
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Lecture Recap

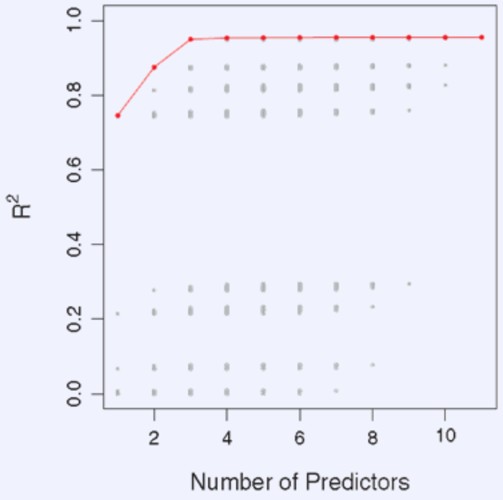
- Subset Selection
  - Only use a subset of the variables in the model. This reduces the flexibility in the model, but a small subset of the coefficients makes the model more interpretable.
  - Find the best model for every possible subset of predictors. There are 2^p such models.
  - However one can also iteratively append or eliminate features greedily. By selecting the predictor which improve the performance most or eliminating feature that reduce performance by least.

Subset Selection

One-standard-error rule:
Choose the simplest model within one standard error of the best model



Best subset selection on the Credit data
Training error measured via RSS



Best subset selection on the Credit data
Training error measured via R^2

Lecture Recap

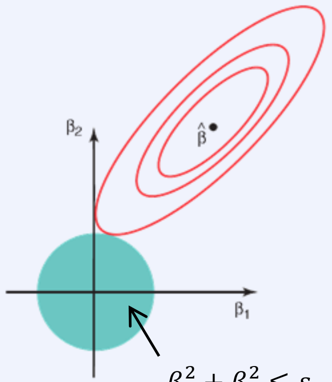
- Shrinkage Methods
  - Penalize models with large or with many non-zero coefficients. The tuning parameter λ adjusts the relative weight of fit and penalty
  - Ridge regression penalizes models that are complex in terms of having large coefficients. While Lasso regression yields naturally sparse models.

Intuition Ridge and Lasso

- Ridge Regression

minimize $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$ such that $\sum_{j=1}^{p}\beta_j^2 \leq s$

- objective defines a circle in coefficient space
- this generalizes to more dimensions

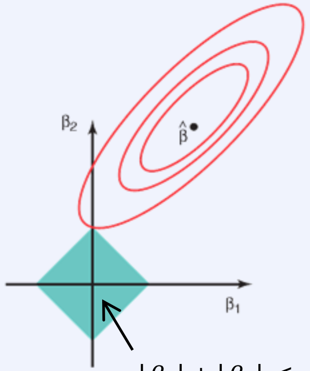


$\beta_1^2 + \beta_2^2 \leq s$

- Lasso

$$\text{minimize} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{such that} \sum_{j=1}^{p}|\beta_j| \leq s$$

- objective defines a diamond in coefficient space
- this generalizes to more dimensions



$|\beta_1| + |\beta_2| \leq s$

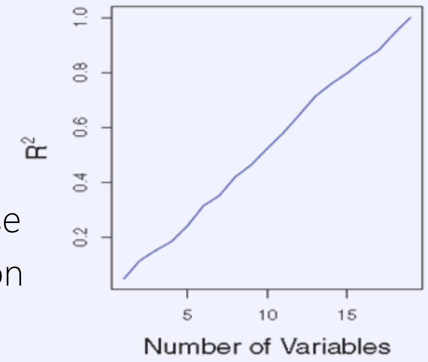
Lecture Recap

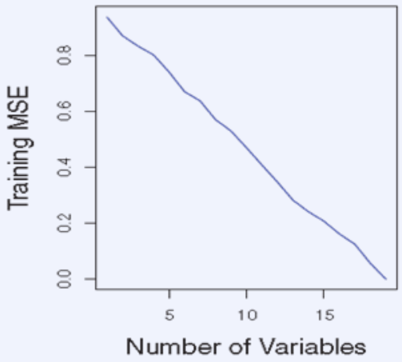
- High Dimensional Data
  - In high dimensions, methods like least squares suggest a perfect fit, but are too flexible and overfit

What Goes Wrong in High-Dimensions

Simulated example

- least-squares regression
- 20 observations
- 1 to 20 features, all completely unrelated to the response
- there is nothing to learn, but nevertheless the correlation rapidly becomes ideal the more features we include
- the training error reduces to zero





What Goes Wrong in High-Dimensions

Simulated example

- least-squares regression
- 20 observations
- 1 to 20 features, all completely unrelated to the response
- there is nothing to learn, but nevertheless the correlation rapidly becomes ideal the more features we include
- the training error reduces to zero
- the test error points very simple models out as the best
- simple model selection techniques like C_p, AIC, BIC do not work well in high-dimensional settings
- adjusted R^2 often approaches 1 and cannot be used either