40.016: Analytics Edge
Week 5 Lecture 2

Model assessment and Model selection: Cross validation and LASSO

Term 6, 2020

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Outline

- Model assessment and Model selection
- Bias-Variance trade-off
- Subset Selection

- Cross validation
- LASSO

Model assessment and Model selection

- We use recent ideas in regression and classification.
- Mostly developed for large data sets with many predictors.
- GOAL – prediction accuracy
  – model interpretability

Bias-Variance trade-off

- Recall the linear regression model fitting problem. The true model is:

$$Y = f(X) + \epsilon$$

where ϵ is a random error term with mean 0 and variance σ^2.

- Using least squares minimization on training data we find predictor $\hat{f}$ for f.
- (X_0, Y_0): (test) data point.

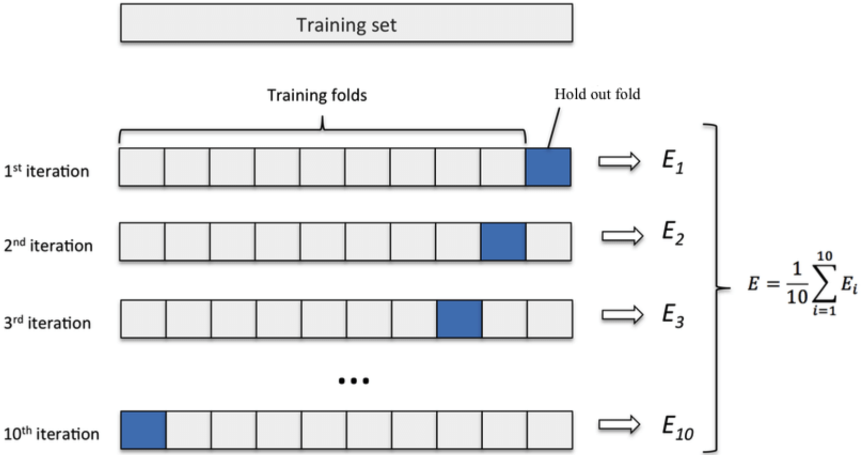
$$\begin{aligned}\text{Test MSE} = \mathbb{E}(Y_0 - \hat{f}(X_0))^2 &= \text{Var}(\hat{f}(X_0)) + \mathbb{E}\left[(f(X_0) - \hat{f}(X_0))^2\right] + \sigma^2 \\ &= \text{Variance of estimator} \\ &\quad + \text{Squared Bias} \\ &\quad\quad + \text{Variance of error term (irreducible errror).}\end{aligned}$$

- Complex model: typically high variance and low bias.
- Simple model: low variance but high bias.

Cross-Validation

- Model assessment technique.
- A model is considered good if it has a low *test set error (TEST MSE)* .
- We often do not have a large test set to validate our model.
- One method of model assessment here is *Cross Validation*.
  - Validation set approach
  - Leave one out cross validation
  - k-fold cross validation.

k-fold cross validation



LASSO

TWO OBJECTIVES:

- Minimize sum of squared errors in the training set.

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 .$$

- Penalize complexity for the model. Minimize $\sum_{i=0}^{p} |\beta_i|$.

LASSO

- LASSO: Least absolute shrinkage and selection operator.
- For a tuning parameter $\lambda \geq 0$:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Balance data fit (first term) with model complexity (second term)

1. When $\lambda = 0$, LASSO reduces to standard linear regression.
2. When $\lambda \uparrow \infty$, the second term dominates and LASSO will make all the beta coefficients for the predictor variables go to zero.

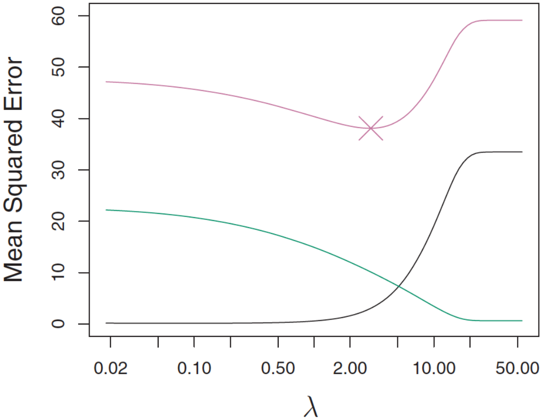
LASSO

1. Proposed in the paper *Regression Shrinkage and Selection via the Lasso*, JRSS B, 1996, by Robert Tibshirani.
2. Around 36000 citations as of October 2020.
3. The objective function in LASSO is convex and tries to roughly promote sparsity.
4. Advantage of LASSO is that since it is convex, the local optimum is the global optimum.
5. Unfortunately, objective function is not differentiable unlike standard linear regression. But there are efficient ways to solve the problem to optimality.

Choice of λ

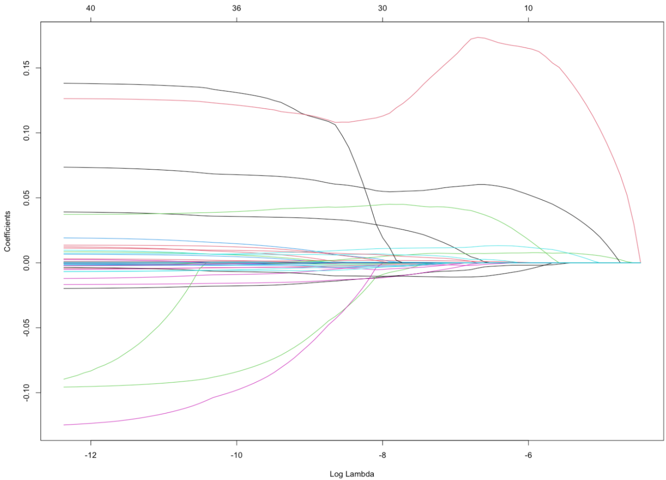
1. Use a grid of possible values and compute the cross-validation error for each value of λ.
2. Choose the λ with the smallest cross-validation error.
3. Finally refit the final model using all the observations for the selected value of λ.

LASSO



- Black line: Squared Bias
- Green line: Variance
- Purple line: Test MSE

LASSO


40
36
30
10
0.15
0.10
0.05
0.00
-0.05
-0.10
Coefficients
-12
-10
-8
-6
Log Lambda

Alternatives to LASSO

- LASSO:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Ridge Regression:

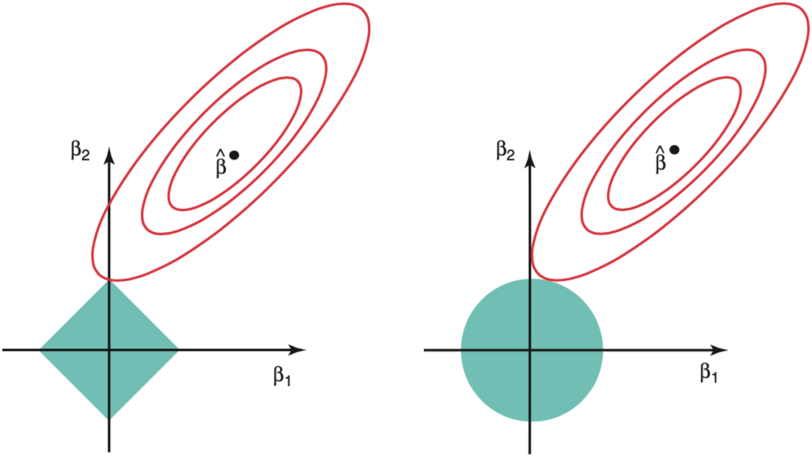
$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

- Elastic Net:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2.$$

– combine ridge regression and LASSO penalty.

LASSO vs Ridge regression



Econometrics: Cross-country growth regression

- Understand factors (economic, political, social) that affect rate of economic growth.
- For example: GDP, degree of capitalism, population growth, equipment investment.
- Robert Barro (1991): Growth rate $\uparrow$ School enrollment rate
  $\downarrow$ Real per capita GDP (1960 level)
- Many such variables have been proposed. Little guidance from economic theory on choice.
- Why not use subset selection from linear regression.
- We use dataset from *I just ran two million regressions* by Sala-I-Martin and *Model uncertainty in cross country growth regression* by Fernandez et. al.
- 41 possible explanatory variables with 72 countries.
- Note that if you try all 2^{41} possible combinations, it leads to around 2 trillion possibilities.