Statistical Learning III (12-11-2015)

11.15 - 13.00, 13.45 - 15.30

1. Model selection intro
2. Best-subset selection
3. Ridge regression
4. Lasso
5. Comparison of ridge and lasso

1. Model Selection intro

Suppose many features p. Maybe even $p >> N$!

Often not satisfied with least squares, because we want:

1. better prediction accuracy:

   (least squares often has low variance, but high variance)

2. better interpretability:

   if many features, which ones are most important
   (e.g. if we have 20 000 genes, want to know which ones are (most) relevant for prediction)

2. Best-subset Selection

Want to use only m most important features.
Select $m \leq p$ features as follows:

1. Run least squares for each of the $\binom{p}{m} \approx \left(\frac{p}{m}\right)^m$ subsets of m features.
2. Choose the subset such that $RSS(\hat{\beta})$ is smallest.

Different ways to choose m. E.g. use cross-validation.

Problems:

* Computation: exponentially many subsets in m
* Variance: discrete choice of subset; if multiple subsets approximately equally good, then which one of those is chosen varies a lot if we would sample a training set multiple times.

Computable approximations:

* Forward stepwise: start with no variables
  greedily choose one variable to add
  until we have m variables
* Backward stepwise: start with all variables,
  greedily choose one variable to remove
  until we have m left.

3. Ridge regression

a) Definition

We set the first feature to 1: $x = \begin{pmatrix} 1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{pmatrix}$

ridge regression: $\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \; RSS(\beta) + \lambda \sum_{j=2}^{p} \beta_j^2$

$$= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_1 - \sum_{j=2}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=2}^{p} \beta_j^2$$

alternative formulation: If $\hat{\beta}$ is unique:

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \; RSS(\beta)$$

$$\text{subject to } \sum_{j=2}^{p} \beta_j^2 \leq t$$

One-to-one relation between λ and t

b) Motivation

* More stable / less variance than least squares for highly correlated features

Example: two features the same: $x_2 = x_3$

Then $RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_1 - \sum_{j=2}^{p} x_{ij} \beta_j \right)^2$ the same for

1.] $\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)$ and

2.] $\beta = (\beta_1, \beta_2 + a, \beta_3 - a, \ldots, \beta_p)$ for any a.

so least squares does not distinguish

ridge regression prefers equal weights (situation 2)

For highly correlated features it makes the weights approximately equal

* $\hat{\beta}_{ridge}$ is always uniquely defined for $\lambda > 0$

Because $RSS(\beta) + \lambda \sum_{j=2}^{p} \beta_j^2$ is strictly convex.

Warning: need to normalize features
- If we make feature x_j 10 times as small, then least squares makes $\hat{\beta}_j$ 10 times as big and its predictions don't change.
- This does not hold for ridge regression, so need to center and rescale features to standard range

c) Computation

- Can interpret as least squares with fake training data.

$$\begin{pmatrix} y_{N+1} \\ x_{N+1} \end{pmatrix}, \dots, \begin{pmatrix} y_{N+p-1} \\ x_{N+p-1} \end{pmatrix}$$

$$y_{N+j} = 0 \qquad x_{N+j} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \sqrt{\lambda} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow (j+1)\text{-th position} \qquad \text{for } j = 1, \dots, p-1$$

$$(y_{N+j} - x_{N+j}^T \beta)^2 = \lambda \beta_{j+1}^2$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{N1} & & x_{Np} \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

with fake training data:

$$\bar{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{N1} & & x_{Np} \\ 0 & \sqrt{\lambda} & 0 \; 0 \cdots \\ 0 & 0 & \sqrt{\lambda} \; 0 \cdots \\ 0 & 0 & 0 \; \sqrt{\lambda} \cdots \\ 0 & 0 & 0 \; 0 \; \ddots \; \sqrt{\lambda} \end{pmatrix} \qquad \bar{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y$$

$$\hat{\beta}_{ridge} = (\bar{X}^T \bar{X})^{-1} \bar{X}^T \bar{y}$$

$$= \left(X^T X + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & \ddots & \\ 0 & 0 & 0 & & \lambda \end{pmatrix}\right)^{-1} Xy$$

$$= (X^T X + \lambda A)^{-1} Xy \qquad A = \begin{pmatrix} 0 & & & \\ & 1 & & 0 \\ & & 1 & \\ & 0 & & \ddots \\ & & & & 1 \end{pmatrix}$$

Different from book, because deals with intercept by pre-processing the data, but is equivalent.

d) Comparison to Least-squares ← if time

$$\hat{\beta}_{ridge} = (X^T X + \lambda A)^{-1} X^T X \cdot \hat{\beta}_{LS}$$

Suppose features orthogonal: $x_j^T x_k = 0$ for $j \neq k$

$$(X^T X + \lambda A)^{-1} X^T X = \begin{pmatrix} \frac{1}{x_1^T x_1} & & 0 \\ & \frac{1}{x_2^T x_2 + \lambda} & \\ & & \ddots \\ 0 & & \frac{1}{x_p^T x_p + \lambda} \end{pmatrix} \begin{pmatrix} x_1^T x_1 & & \\ & \ddots & \\ & & x_p^T x_p \end{pmatrix}$$

$$\hat{\beta}_{ridge,1} = \hat{\beta}_{LS,1}$$

$$\hat{\beta}_{ridge,j} = \frac{x_j^T x_j}{x_j^T x_j + \lambda} \hat{\beta}_{LS,j} \quad \text{for } j = 2, \ldots, p$$

← shrinkage of least squares estimate

Suppose features centered: $\frac{1}{N} \sum_{i=1}^{N} x_{ij} = 0$

Then $x_j^T x_j = \sum_{i=1}^{N} x_{ij}^2$ is empirical variance in feature j.

Hence Effect of λ is larger when $x_j^T x_j$ = variance in feature j is small

So ridge shrinks most along direction in which empirical variance in features is smallest.

this still holds when features not orthogonal: Figure 3.9

Implicit assumption: Y varies most in directions of high variance

4. Lasso

a] Definition

$$\hat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}} \; RSS(\beta) + \lambda \sum_{j=2}^{p} |\beta_j|$$

If $\hat{\beta}_{LS}$ unique, then

$$\hat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}} \; RSS(\beta)$$

$$\text{subject to } \sum_{j=2}^{p} |\beta_j| \leq t$$

One-to-one relation: $t \leftrightarrow \lambda$

b] Motivation

* Sets many weights β_j to 0 to find only most important features: good for prediction and interpretation

* $p \gg N$ allowed

NB Need to normalize features

c] Computation

- Exist multiple efficient algorithms
- LARS computes solutions for all values of t in one pass

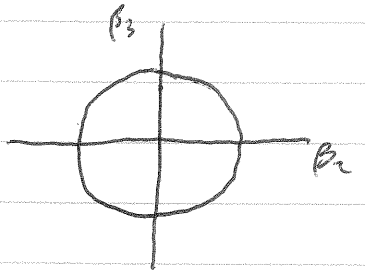
5. Comparison of Ridge, Lasso, Best-subset Selection

a) Ridge, Lasso vs Best-subset

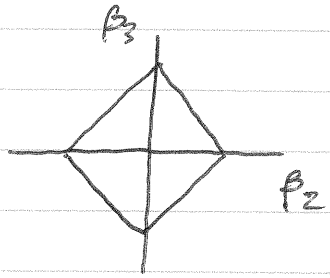
↑ small changes in data cause small changes in estimate, so less variance than best-subset.

b) Ridge vs Lasso

Ridge
$\sum_{j=2}^{p} \beta_j^2 \leq t$

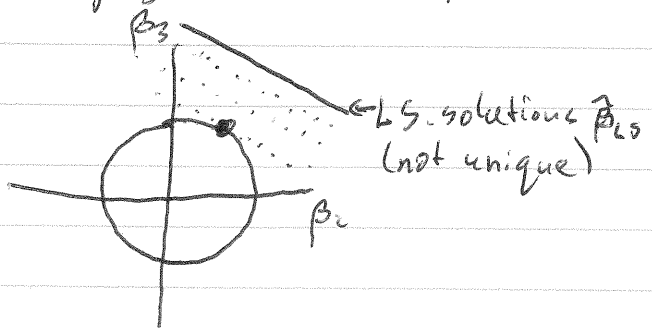


Lasso
$\sum_{j=1}^{p} |\beta_j| \leq t$

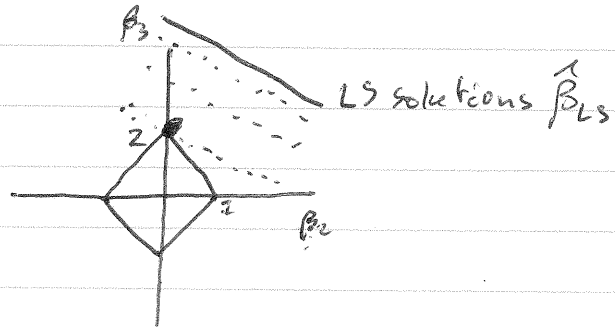


* Figure 3.11

* For highly correlated features:



β_2 and β_3 get about equal weight



minor noise determines which of corners 1 and 2 is chosen: one variable gets all weight, the other 0.

6. Homework 1

* Goal: try out linear regression methods on real data: predicting housing prices

* One of the features is racist
  Bad? Depends on what you do with predictions.

* Statistician's responsibility goes beyond technical aspects. Management does not know details of the methods used!