Penalised regression

Krydsvalidering – Ridge regression and LASSO

Introduction to Data Science

Torben Tvedebrink
tvede@math.aau.dk

Department of Mathematical Sciences

AALBORG UNIVERSITY
DENMARK

Version: 24/11/2017 11:41

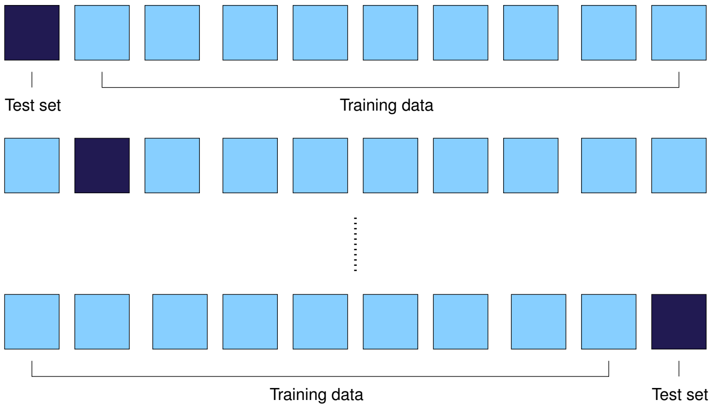
Krydsvalidering

En effektiv og simpel og effektiv måde til at estimere *generaliseringsfejlen* for en statistisk model/metode er vha. **krydsvalidering**.

K-fold krydsvalidering går ud på at imitere processen hvor vi har adgang til nye testdata, ved at vi opdeler data i K dele og successivt bruger de $K-1$ dele som træningsdata og den resterende del til testdata.

Torben Tvedebrink
tvede@math.aau.dk

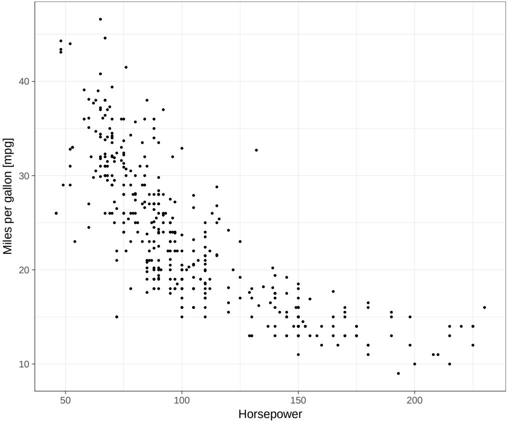
10-fold krydsvalidering



Torben Tvedebrink
tvede@math.aau.dk

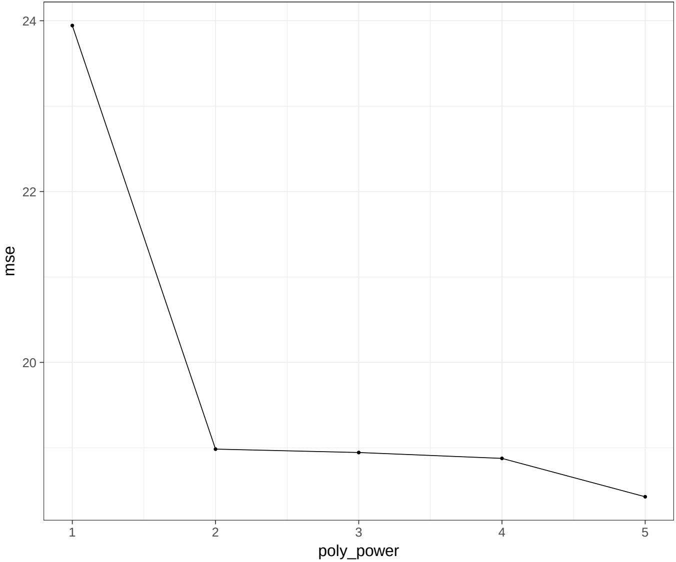
Eksempel

Vi kan fx. lave 10-fold krydsvalidering til at sige noget om passende valg af kompleksitet for en given model.

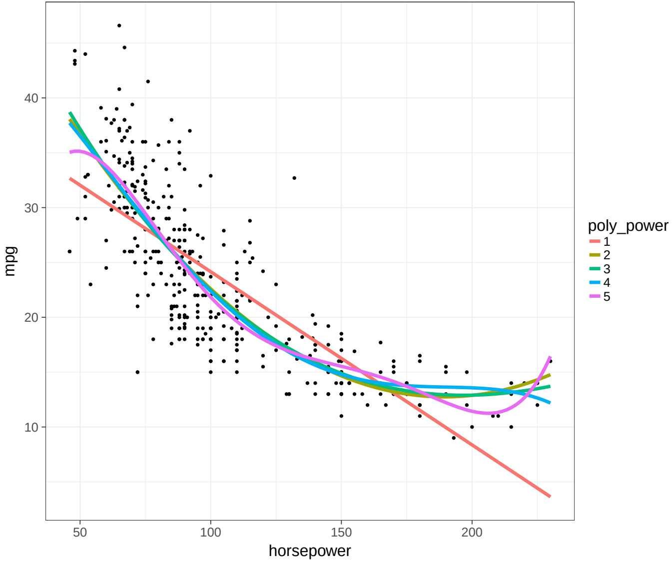


Torben Tvedebrink
tvede@math.aau.dk

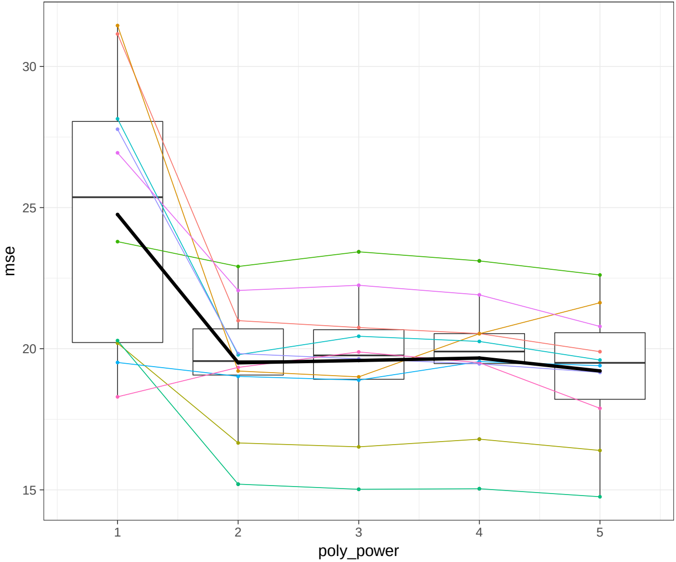
Eksempel



Eksempel



Eksempel



Bootstrap

Torben Tvedebrink
tvede@math.aau.dk

Idéen bag bootstrap minder om krydsvalidering, idet vi benytter re-sampling af vores data til at estimere standard errors og bias af parameter estimater.

Normalt noteres en bootstrap sample med $\boldsymbol{x}^* = (x_1^*, \dots, x_n^*)$, hvor x_i^* er trukket med tilbagelægning fra $x_1, \dots, x_n$.

Baseret på $\boldsymbol{x}^*$ kan vi således estimere de ukendte parametre θ og opnå et estimat $\hat{\theta}^*$. Dette kan vi gøre N gange, hvorved vi har $\hat{\theta}_1^*, \dots, \hat{\theta}_N^*$ estimater af θ, hver baseret et bootstrap sample $\boldsymbol{x}_i^*$, $i = 1, \dots, N$.

Standard error og bias

Vi kan således sige noget om variabiliteten af $\hat{\theta}^*$ omkring $\hat{\theta}$, idet begge er funktioner af de observerede data (og re-samplinger af disse).

Ved en hver form for inferens er vi interesserede i at sige noget variationen af $\hat{\theta}$ omkring den sande værdi θ.

Ved bootstrap kan vi tilgå denne information ved at estimere $\mathrm{sd}(\hat{\theta})$ ved $(N-1)^{-1}\sum_{j=1}^{N}(\hat{\theta}-\hat{\theta}_j^*)^2$.

Torben Tvedebrink
tvede@math.aau.dk

Bootstrap interval

Ud fra vores bootstrap estimatater $\theta_1^*, \ldots, \theta_N^*$ kan vi forme et *bootstrap interval* for estimatet af θ.

Fx. hvis $N = 100$ er $[\theta_{(5)}^*, \theta_{(95)}^*]$ et approksimativt 90%-konfidens interval for θ, hvor $\theta_{(1)}^* \leq \theta_{(2)}^* \leq \cdots \leq \theta_{(N)}^*$ er ordnet efter størrelse.

Torben Tvedebrink
tvede@math.aau.dk

Eksempel

Torben Tvedebrink
tvede@math.aau.dk

Antag at X og Y repræsenterer to afkastet fra to investerings aktiver. Vi ønsker at minimere investerings risikoen, hvilket svarer til at minimere variansen af $\alpha X + (1-\alpha)Y$, hvor α angiver andelen investeret i X.

Vi kan vise (*opgave*) at det optimale α angives som

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

hvor $\sigma_X^2 = \mathbb{V}(X)$, $\sigma_Y^2 = \mathbb{V}(Y)$ og $\sigma_{XY} = \mathbb{C}(X, Y)$.

Department of Mathematical Sciences

Eksempel (fortsat)

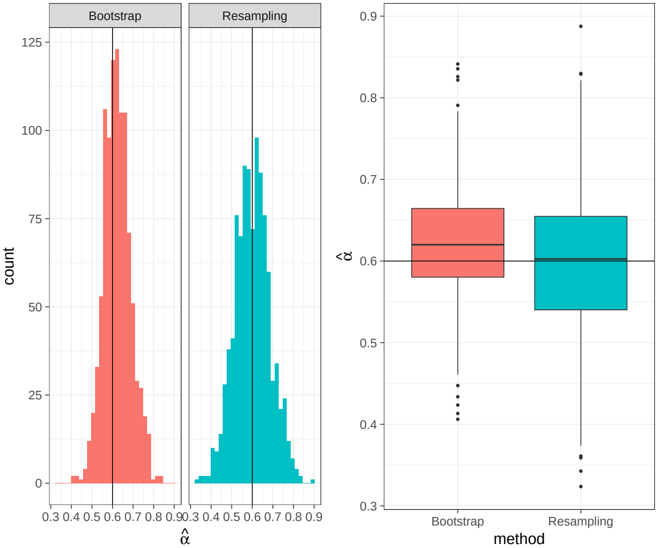
Generelt er σ_X^2, σ_Y^2 og σ_{XY} ukendte og estimeres derfor fra data.

Lad os antage at vi kender de sande værdier. Vi kan således simulere data fra den sande fordeling og se variabiliteten i estimatet for α.

I det følgende simulerer vi 1000 datasæt med 100 observationer i hver. Den første simulation benytter vi efterfølgende til at lave et bootstrap af (ligeledes 1000 bootstrap samples).

Torben Tvedebrink
tvede@math.aau.dk

Simuleret vs. bootstrap



Bet on sparsity

Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

Torben Tvedebrink
tvede@math.aau.dk

Our point of departure

In linear regression we assume that the ith response, y_i, can be modelled using a linear relationship between some covariates and the response with an additive error term with constant variance

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

Our point of departure

In linear regression we assume that the ith response, y_i, can be modelled using a linear relationship between some covariates and the response with an additive error term with constant variance

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

If we have observations, $i = 1, \dots, n > p$, we have that the least squares estimator for β_0 and $\beta = (\beta_1, \dots, \beta_p)$ is given by

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0, \beta} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2$$

Least squares
On a *budget*

Imagine that we only had a limited *budget* of regression coefficients, t, such that the sum $\sum_{j=1}^{p} h(\beta_j)$ was restricted by t, then the solution should obey this constraint

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \quad \text{such that} \quad \sum_{j=1}^{p} h(\beta_j) \leq t$$

Torben Tvedebrink
tvede@math.aau.dk

Least squares
On a *budget*

Imagine that we only had a limited *budget* of regression coefficients, t, such that the sum $\sum_{j=1}^{p} h(\beta_j)$ was restricted by t, then the solution should obey this constraint

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad \text{such that} \quad \sum_{j=1}^{p} h(\beta_j) \leq t$$

For

- $h(\beta_j) = |\beta_j|$ we term the regression problem the *LASSO*, and
- $h(\beta_j) = \beta_j^2$ we refer to the problem as *ridge regression*.

Torben Tvedebrink
tvede@math.aau.dk

Reasons for abandoning least squares

- The *prediction accuracy* can sometimes be improved because even though least squares has zero bias, its high variance may cause bad prediction ability. Hence, shrinking some coefficients, or setting the *noisy terms* to zero, may improve the accuracy.

Torben Tvedebrink
tvede@math.aau.dk

Reasons for abandoning least squares

- The *prediction accuracy* can sometimes be improved because even though least squares has zero bias, its high variance may cause bad prediction ability. Hence, shrinking some coefficients, or setting the *noisy terms* to zero, may improve the accuracy.
- The second reason is *interpretation*. The fewer terms to interpret the easier it gets.

Reasons for abandoning least squares

- The *prediction accuracy* can sometimes be improved because even though least squares has zero bias, its high variance may cause bad prediction ability. Hence, shrinking some coefficients, or setting the *noisy terms* to zero, may improve the accuracy.
- The second reason is *interpretation*. The fewer terms to interpret the easier it gets.
- The third reason being that it fails for *wide* data, i.e. data for which $p \gg n$

Torben Tvedebrink
tvede@math.aau.dk

Standardisation of $\mathbf{X}$

As the *numerical value* of coefficients is sensitive to the scale of the covariates, it is typically preferred to standardise the $\mathbf{X}$ matrix before estimating the coefficients. That is,

$$\sum_{i=1}^{n} x_{ij} = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_{ij}^2 = n$$

Torben Tvedebrink
tvede@math.aau.dk

Standardisation of $\mathbf{X}$
and centering of $\boldsymbol{y}$

As the *numerical value* of coefficients is sensitive to the scale of the covariates, it is typically preferred to standardise the $\mathbf{X}$ matrix before estimating the coefficients. That is,

$$\sum_{i=1}^{n} x_{ij} = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_{ij}^2 = n$$

And in order to discard the intercept, β_0, from the regularisation in the case of linear regression we center the response

$$\sum_{i=1}^{n} y_i = 0$$

Torben Tvedebrink
tvede@math.aau.dk

The *wide* data problem

Torben Tvedebrink
tvede@math.aau.dk

In the case where $p \gg n$, the least squares estimator is undefined as $(\mathbf{X}^\top \mathbf{X})$ isn't invertible because $\mathbf{X}$ is not of full rank. Hence, $\hat{\beta}^{\text{ols}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$ cannot be evaluated.

The *wide* data problem

In the case where $p \gg n$, the least squares estimator is undefined as $(\mathbf{X}^\top \mathbf{X})$ isn't invertible because $\mathbf{X}$ is not of full rank. Hence, $\hat{\beta}^{\text{ols}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y}$ cannot be evaluated.

A solution to this is to add an invertible matrix to $\mathbf{X}^\top \mathbf{X}$ to obtain an invertible matrix. The simplest such candidate is $\lambda \mathbf{I}_p$, for some positive $\lambda \in \mathbb{R}$:

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^\top \boldsymbol{y},$$

which is what is referred to as the ridge regression estimator.

Torben Tvedebrink
tvede@math.aau.dk

Ridge regression

For the least squares regression problem with a budget on the squared entries of β we have

$$\min_{\beta} \sum_{i=1}^{2} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 \quad \text{such that} \quad \sum_{j=1}^{p} \beta_j^2 \leq t.$$

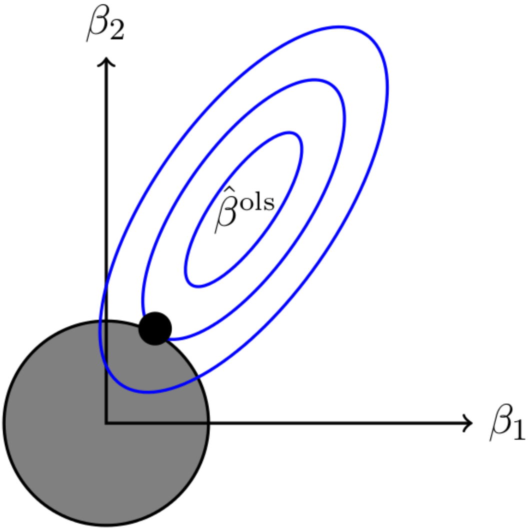
This can also be stated as

$$\min_{\beta} \sum_{i=1}^{2} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Torben Tvedebrink
tvede@math.aau.dk

Visual representation of $\hat{\beta}^{\text{ridge}}$

Compared to $\hat{\beta}^{\text{ols}}$ (in two dimensions)



Torben Tvedebrink
tvede@math.aau.dk

Now, what happens if we instead of using a squared penalty, β_j^2, uses the absolute penalty, $|\beta|$?

Well – we obtain the LASSO

$$\min_{\beta} \sum_{i=1}^{2} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \quad \text{such that} \quad \sum_{j=1}^{p} |\beta_j| 2 \leq t.$$

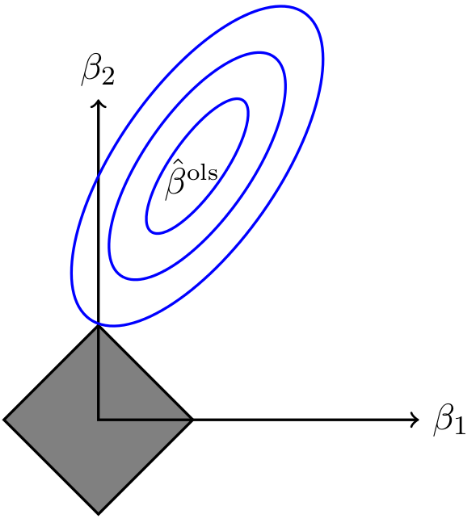
and again an equivalent form

$$\min_{\beta} \sum_{i=1}^{2} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \lambda \sum_{j=1}^{p} |\beta_j|.$$

Torben Tvedebrink
tvede@math.aau.dk

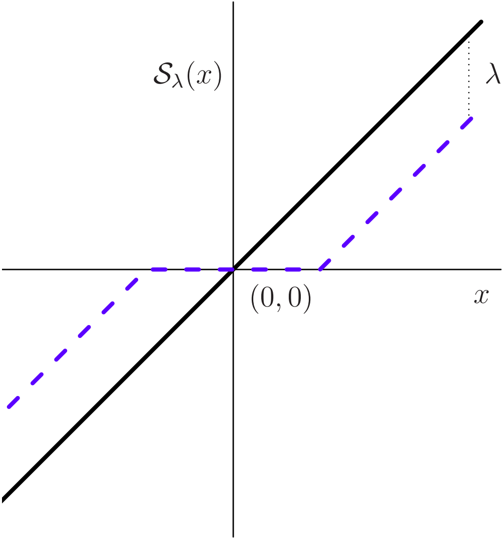
Visual representation of $\hat{\beta}^{\text{lasso}}$

Compared to $\hat{\beta}^{\text{ols}}$ (in two dimensions)



Soft thresholding

$\mathcal{S}_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$



Elastic Net

The best from two worlds?

Torben Tvedebrink
tvede@math.aau.dk

A downside with the Lasso is that it may have difficulties when several variables are collinear, such that linear combinations of them are hard to distinguish.

In such a case the Ridge Regression is better as it will typically form an average of the variables. Hence, for stable selection of variables in this case Ridge Regression may be preferred.

However, Ridge Regression seldom sets any parameters to zero, i.e. no variable selection which is what we would like in the end...

The solution to the problem is Elastic Net, which incorporates both the Lasso and Ridge penalties in a convex way:

$$\min_{\beta} \sum_{i=1}^{2} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 + \lambda \sum_{j=1}^{p} \{\alpha|\beta_j| + (1-\alpha)\beta_j^2\},$$

where α is yet another tuning parameter deciding the amount of Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$) penalty that goes into the solution.

Both α and λ are selected based on cross-validation.

Torben Tvedebrink
tvede@math.aau.dk

Department of Mathematical Sciences

Elastic Net
The best from two worlds?

Torben Tvedebrink
tvede@math.aau.dk

In the Figure below we see the three types of regularisation discussed above. The shape of the Elastic Net solution area depends on α - the closer to 1 the more square it is, and the closer to 0 the more spherical.