Lerntext zu LASSO

Peter von Rohr

2016-05-17

Dokumentenstatus

version	author	date	status	project
0.0.901	peter	2016-05-09	Initialisierung	ASMAS

Abkürzungen

Abk	Bedeutung
LASSO	Least Absolute Shrinkage and Selection Operator
RSS	Restsummenquadrate (Residual Sums of Squares

Contents

Erklärung

Dieses Dokument gibt einerseits eine Zusammenfassung zum Thema LASSO und ist andererseits als Lerntext organisiert. Das heisst, nach jedem Abschnitt wird eine Kontrollfrage gestellt. Kann diese Frage beantwortet werden, können wir im Text weiterfahren, sonst empfehle ich, den der Frage vorangegangenen Abschnitt noch einmal zu lesen. Der Inhalt basiert im Wesentlichen auf Kapitel 6 von Gareth et al. (2013).

Einführung

Das lineare Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \tag{1}$$

wird verwendet um Zusammenhänge zwischen einer Beobachtung y_i (Zielgrösse) und erklärenden Variablen $x_{i1}, x_{i2}, ..., x_{ip}$ zu beschreiben. Zusätzlich zu einer Beobachtung y_i haben wir auch noch p Werte von erklärenden Variablen. Somit resultiert der folgenden Vektor an Informationen für das Ergeignis i:

$$(x_{i1}, x_{i2}, ..., x_{ip}, y_i)$$

Die $(p+1)$ Werte $\beta_0, ..., \beta_p$ und ϵ_i sind unbekannt. Es wird angenommen, dass die Werte der erklärenden Variablen $(x_{i1}, x_{i2}, ..., x_{ip})$ exakt, d.h. ohne Messfehler oder andere Ungenauigkeiten, bekannt sind. Für einen Datensatz mit n Beobachtungen werden die resultierenden n Gleichungen vorzugsweise in Matrix-Vektor-Schreibweise notiert.

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{2}$$

Kontrollfrage 1: Teilen Sie die nachfolgenden Komponenten in die zwei Kategorien bekannt oder unbekannt ein

Komponenten	bekannt	unbekannt
y_i		
$x_{i1}, x_{i2}, ..., x_{ip}$		
$\beta_0, ..., \beta_p$		
ϵ_i		

Stochastische Restkomponente

Die n unbekannten Resteffekte im Vektor ϵ werden als zufällige Effekte modelliert, wobei angenommen wird, dass sich diese Resteffekte im Mittel aufheben, d.h., dass deren Erwartungswert $E(\epsilon) = \mathbf{0}$ ist. Die Streuung der Resteffekte wird im Standardmodell als konstant angenommen. Für die Covarianz des Vektors der Resteffekte bedeutet das, dass $var(\epsilon) = \mathbf{I} * \sigma^2$ ist. Die Varianzkomponente σ^2 ist neben den Koeffizienten im Vektor β ein weiterer unbekannter Parameter, welcher von den Daten geschätzt werden muss.

Parameterschätzung

Unter der Annahme, dass die Matrix $\mathbf{X}$ vollen Kolonnenrang hat, d.h. die Anzahl Beobachtungen n grösser ist als die Anzahl Parameter (hier $p+1$) lassen sich die unbekannten Parameter β mit **Least Squares** schätzen. Der Least Squares Schätzer $\hat{\beta}$ für β wird berechnet aus

$$\hat{\beta} = argmin_{\beta}||\mathbf{y} - \mathbf{X}\beta||^2 \tag{3}$$

wobei $||.||$ für die Euklidsche Norm (Länge) im n-dimensionalen Raum steht. Wird das Minimierungsproblem in Gleichung (3) aufgelöst, dann resultiert der folgende Ausdruck für $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

Betrachten wir den Ausdruck in Gleichung (4) wird klar, weshalb die Matrix $\mathbf{X}$ vollen Kolonnenrang haben muss, da nur so die Inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ berechnet werden kann.

Kontrollfrage 2:

a. Welche Anforderung bezüglich des Ranges der Matrix $\mathbf{X}$ besteht?
b. Aus welchem Grund besteht diese Anforderung aus 2a?

Alternativen zu Least Squares

Das lineare Modell (1) erweist sich in der Praxis als sehr brauchbar. Mit der Least Squares-Technik besteht auch eine einfache und sehr gut etablierte Methode zur Parameterschätzung. In kürzerer Vergangenheit auch mit dem Aufkommen des Phänomes von "Big Data", welches das systematische Sammeln von grossen Datenmengen ermöglicht, treten häufiger Probleme auf, bei welchen die im einleitenden Abschnitt aufgestellte Bedingung an Least Squres, dass nämlich $n > p$ gelten muss, nicht zutrifft.

Da wir die positiven Eigenschaften des linearen Modells gerne beibehalten möchten, wurde nach Alternativen zu Least Squres gesucht. Diese möglichen Alternativen können in drei Kategorien eingeteilt werden.

1. **Subset Selektion**: Aus den p erklärenden Variablen wird ein Subset von "relevanten" Variablen ausgewählt. Alle anderen Variablen werden ignoriert. Die relevanten Variablen werden oft aufgrund der Signifikanz des geschätzten Regressionskoeffizienten β_j identifiziert.
2. Regularisierung.(Shrinkage): Alle p erklärenden Variablen werden verwendet. Die geschätzten Regressionskoeffizienten werden durch bestimmte Techniken gegen den Nullpunkt "gedrückt". Dieser Prozess wird als Schrumpfung (Shrinkage) bezeichnet. Die so erzeugte Reduktion der Variabilität der Schätzwerte wird als Regularisation bezeichnet.
3. **Dimensionsreduktion**: Die p erklärenden Variablen werden zu m Linearkombinationen reduziert. Diese Reduktion kann mit Techniken, wie Principal Components Analysis oder Faktoranalyse gemacht werden.

Kontrollfrage 3:

a. Wieso brauchen wir Alternativen zu Least Squares?
b. Wie sehen die Alternativen zu Least Squares aus?

Lasso

Es gibt Schätzverfahren, welche mehrere der oben genannten Alternativen zu Least Squares kombinieren. Ein Beispiel dafür ist LASSO. LASSO steht für Least Absolute Shrinkage and Selection Operation und kombiniert "Subset Selection" und Regularisierung. Die Regularisierung wird durch das Hinzufügen eines Terms zu den Rest-Summenquadraten (RSS), welche bei Least Squares minimiert werden. In Gleichung (3) haben wir gesehen, wie RSS verwendet werden zur Berechnung der Least Squares Schätzer

$$\begin{aligned} \hat{\beta}_{LS} &= argmin_\beta ||\mathbf{y} - \mathbf{X}\beta||^2 \\ &= argmin_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \\ &= argmin_\beta RSS \end{aligned} \tag{5}$$

Regularisierung bei LASSO

Bei LASSO wird nun zu RSS ein sogenannter Strafterm (penalty term) hinzugefügt. Dieser Strafterm beträgt $\lambda \sum_{j=1}^{p} |\beta_j|$. Der Term wird deshalb als Strafterm bezeichnet, weil er mit steigender Summe der Absolutbeträge aller β_j immer grösser wird. Diese führt zum gewünschten Effekt der Regularisierung. Das heisst durch das Hinzufügen dieses Strafterms werden die Absolutbeträge und somit die Variabilität der Koeffizientenschätzungen begrenzt, was der eigentliche Sinn und Zweck der Regularisierung ist.

In Formeln ausgedrückt, lauten die geschätzten Regressionskoeffizienten für LASSO, wie folgt:

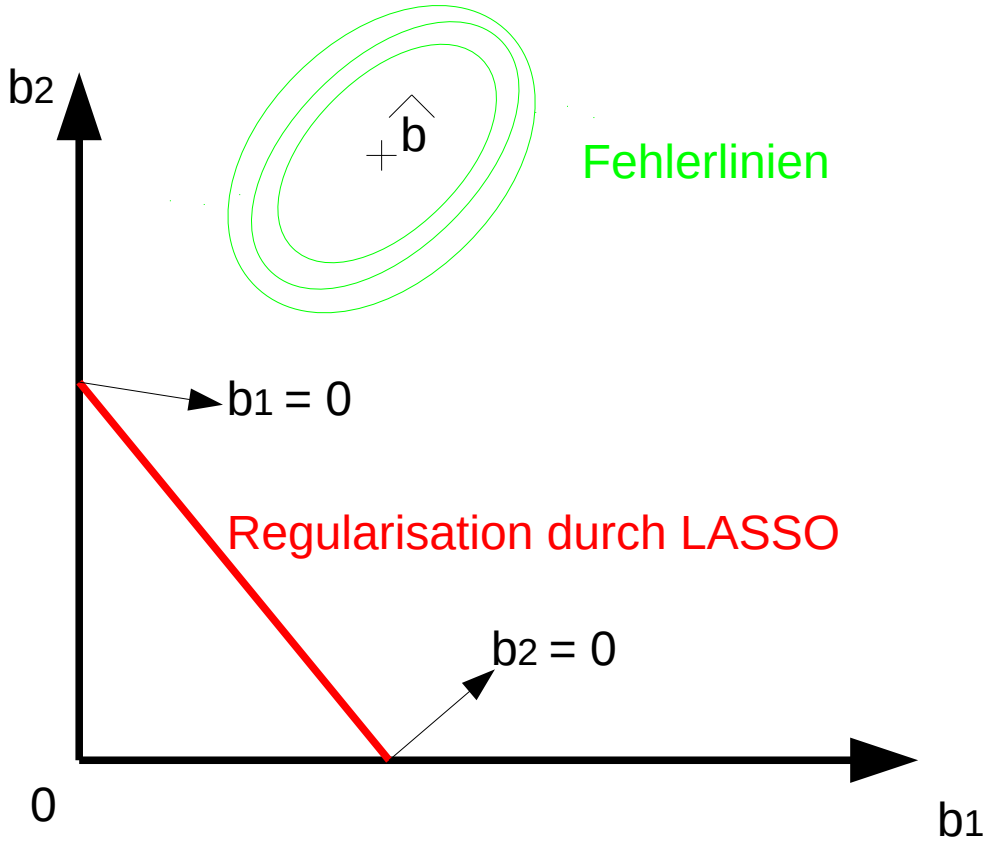
$$\begin{aligned} \hat{\beta}_{LASSO} &= argmin_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\ &= argmin_\beta \left\{ RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \end{aligned} \tag{6}$$

Subset Selection bei LASSO

Wie schon im vorangegangenen Abschnitt beschrieben, dient der Strafterm $\lambda \sum_{j=1}^{p} |\beta_j|$ zur Regularisierung der geschätzten Koeffizienten β_j im linearen Modell. Der Strafterm spielt auch eine entscheidene Rolle bei der Subset Selection. Dadurch, dass der Strafterm die Absolutbeträge der Koeffizienten β_j summiert, werden die Schätzungen von gewissen Koeffizienten explizit auf Null gesetzt. Weshalb dieser Effekt der Subset Selection bei LASSO eintritt kann mit folgender Abbildung (siehe nächste Seite) erklärt werden.

In dieser Abbildung sind nur zwei erklärende Variablen gezeigt und somit ist $p = 2$. Die Koeffizienten zu den erklärenden Variablen werden in der Abbildung mit b und nicht mit β bezeichnet. Unter der Annahme, dass wir unendlich viele Daten hätten, wäre der Schätzer der Koeffizienten b_j mit minimalem Fehler am Punkt, welcher in der Abbildung mit $\hat{b}$ bezeichnet ist. Die grünen Ellipsen um diesen Punkt $\hat{b}$ sind die Linien mit konstantem Fehler. Die rote Linie steht für die Grenze, welche durch den Strafterm aus LASSO entsteht. Das heisst geschätzte Koeffizienten können nur links dieser roten Linie liegen. Da wir den geschätzten Koeffizienten $\hat{b}_j$ einerseits minimalen Fehler erreichen wollen und auf der anderen Seite innerhalb der Regularisierungsgrenzen sein müssen, liegen die besten Schätzer für b_j am Schnittpunkt zwischen den grünen Ellipsen und der roten Linie. Durch den Verlauf der roten Linie ist die Wahrscheinlichkeit, dass sich die grünen Ellipsen und die rote Linie auf einer Koordinatenachse schneiden sehr hoch. Schneiden sich die grünen Ellipsen und die rote

Linie auf einer Koordinatenachse, dann wurde ein Schätzer für einen Koeffizienten b_j auf Null gesetzt und somit haben wir den gewünschten Effekt der Subset Selection erreicht.



Kontrollfrage 4:

a. Wie unterscheidet sich ein LASSO-Schätzer β_{LASSO} von einem Least Squares Schäter β_{LS}?
b. Wie erreichen wir mit LASSO den Effekt der Regularisierung?
c. Wie werden mit LASSO gewissen erklärende Variablen selektioniert?

Bestimmung von λ

Der Strafterm, welcher in Gleichung (6) eingefügt wurde und für die Regularisierung bei LASSO verantwortlich ist, enhält eine Variable λ. Diese Variable bestimmt das Ausmass der Regularisierung und muss als zusätzlicher Parameter aus den Daten bestimmt werden. Für die Bestimmung von λ wird eine sogenannte Kreuzvalidierungsprozedur (cross validation) verwendet. Bei einer Kreuzvalidiuerng werden die Beobachtungen zufällig in ein sogenanntes Trainings-Set und in ein Test-Set unterteilt, wobei das Test-Set meist weniger Beobachtungen enthält als das Trainings-Set. Mit dem Trainings-Set werden dann die Koeffizienten β_j geschätzt. Dann werden für vorher bestimmte Werte von λ die Beobachtungen im Test-Set vorhergesagt. Der Wert von λ, welcher die tiefsten Vorhersagefehler liefert, wird als optimaler Schätzwert von λ betrachtet.

Analyse mit LASSO in R

In diesem Abschnitt wird gezeigt, wie ein Datensatz mit LASSO in R analysiert werden kann. Wir verwenden dazu den `Hitters`- Datensatz aus dem Buch von Gareth et al. (2013). Dieser Datensatz enthält als Zielgrösse das Einkommen von Baseballspielern und zu diesen Spielern noch weitere erklärende Variablen. Der Datensatz ist im R-Package `ISLR` integriert. Für die Analyse werden wir die Funktion `glmnet()` aus dem gleichnamigen R-Package verwenden. Als erstes installieren wir die beiden Packages und ignorieren alle Records, welche fehlende Daten aufweisen.

```
if (!require(ISLR)) {
  install.packages("ISLR")
  require(ISLR)
}
```

```
## Loading required package: ISLR
```

```
if (!require(glmnet)){
  install.packages("glmnet")
  require(glmnet)
}
```

```
## Loading required package: glmnet

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-5
```

```
### # records mit fehlenden Daten ignorieren
data(Hitters)
Hitters <- na.omit(Hitters)
dim(Hitters)
```

```
## [1] 263  20
```

Da wir für die Bestimmung von λ mit Kreuzvalidierung ein Trainings- und ein Test-Set benötigen, bestimmen wir diese durch den Zufallszahlengenerator und der Funktion `sample()`

```
set.seed (1)
train <- sample (c(TRUE ,FALSE), nrow(Hitters), rep=TRUE)
test  <- (! train )
```

Wir verwenden die Funktion `glmnet()` zur Modellierung mit LASSO. Für diese Funktion muss das Modell anders spezifiziert werden als für die Funktion `lm()`. Wir brauchen dazu die Objekte `x` und `y`.

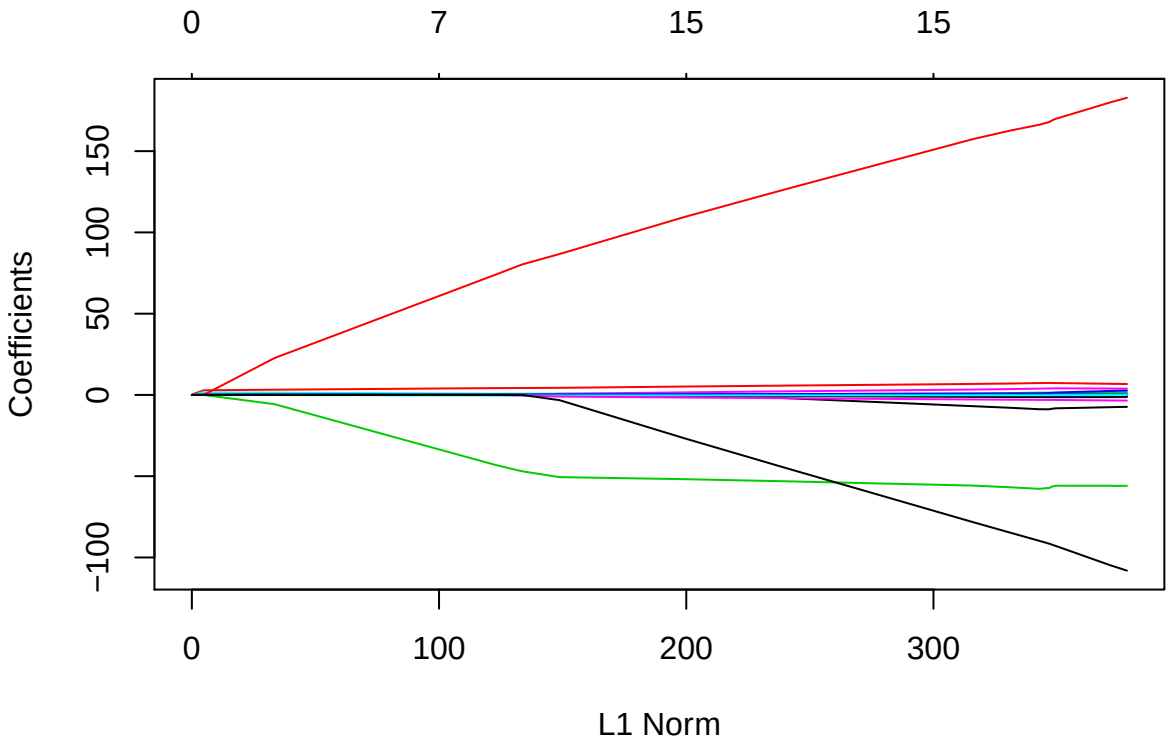
```
x <- model.matrix (Salary ~ ., Hitters)[,-1]
y <- Hitters$Salary
```

Die vorgegebenen Werte für λ werden in der Variablen `grid` abgelegt. Es handelt sich um 100 Werte zwischen 10^10 und 10^{-2}.

```
grid <- 10^ seq (10,-2, length =100)
```

The following statements fits a LASSO model.

```
lasso.mod <- glmnet (x[train ,],y[train],alpha =1, lambda = grid)
plot(lasso.mod)
```



Der Plot zeigt, wie sich der Strafterm für verschiedene Werte (durch Farben codiert) verhält. Nun wollen wir den besten Wert für λ bestimmen. Dies wird durch Kreuzvalidierung gemacht.

```
set.seed (1)
cv.out <- cv.glmnet (x[train ,],y[train],alpha =1)
bestlam <- cv.out$lambda.min
```

Der Anteil an Koeffizienten, welcher durch LASSO null gesetzt wird kann mit folgenden Statements überprüft werden.

```
out <- glmnet(x, y, alpha = 1, lambda = grid)
lasso.coef <- predict(out, type = "coefficients", s=bestlam )[1:20,]
lasso.coef
```

```
##   (Intercept)         AtBat          Hits         HmRun          Runs
##  8.898370e-01 -5.575622e-03  2.007078e+00  0.000000e+00  0.000000e+00
##           RBI         Walks         Years        CAtBat         CHits
##  0.000000e+00  2.268641e+00 -3.428874e-02  0.000000e+00  0.000000e+00
##        CHmRun         CRuns          CRBI        CWalks       LeagueN
##  8.315024e-03  2.102106e-01  4.211554e-01  0.000000e+00  1.695962e+01
##     DivisionW       PutOuts       Assists        Errors    NewLeagueN
## -1.143553e+02  2.343374e-01  0.000000e+00 -6.607899e-01  0.000000e+00
```

Antworten zu den Kontrollfragen

Antwort 1:

Komponenten	bekannt	unbekannt
y_i	x	
$x_{i1}, x_{i2}, ..., x_{ip}$	x	
$\beta_0, ..., \beta_p$		x
ϵ_i		x

Antwort 2:

a. Matrix $\mathbf{X}$ muss vollen Kolonnenrang haben, d.h. $n > p$ sein
b. Nur so ist die Inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ berechenbar

Antwort 3:

a. Least Squares kann nur verwendet werden, wenn $n > p$ gilt, d.h. die Anzahl der Beobachtungen grösser ist als die Anzahl zu schätzender Parameter. Da es aber immer häufiger Anwendungen gibt, bei denen das nicht der Fall ist, brauchen wir Alternativen zu Least Squares.

b. Es wurden drei Alternativen vorgestellt: (1) Subset-Selection, (2) Regularisation und (3) Dimensionsreduktion

Antwort 4:

a. Der Unterschied liegt im Strafterm (penalty-term). Dieser tritt nur bei LASSO auf, nicht aber bei Least-Squares
b. Der Strafterm $\lambda \sum_{j=1}^{p} |\beta_j|$ wächst mit grösseren Absolutbeträgen der Schätzungen. Somit wird die Variabilität der Koeffizientenschätzungen eingegrenzt.
c. Die Subset Selection wird durch die Art des Strafterms erreicht. Da beim Strafterm die Absolutbeiträge der Koeffizientenschätzungen summiert werden, werden gewisse Schätzungen von Koeffizienten explizit auf Null gesetzt. Die erklärenden Variablen mit Koeffizientenschätzungen von 0 werden im linearen Modell nicht berücksichtigt. Somit kommt es zu einer Selektion der erklärenden Variablen.

References

Gareth, J, D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*. Edited by Springer.