Elastic Net «The Oracle»
Poor Man's Introduction To Penalized Regression

Sinan İyisoy

Plan

- OLS Regression
- Ridge Regression
- Lasso
- Bridge Regression
- Elastic Net
- Uygulama

L_q Norm Ailesi

- n boyutlu bir $x = (x_1, x_2, \ldots x_n)$ vektörü için

L_1 normu $\|x\|_1 = |x_1|+|x_2|+..+|x_n|,$

L_2 normu $\|x\|_2 = (x_1{}^2 + x_2{}^2+..+x_n{}^2)^{1/2},$

L_q normu $\|x\|_q = (x_1{}^q + x_2{}^q+..+x_n{}^q)^{1/q}$

şeklinde tanımlanır.

OLS Regression

En küçük kareler regresyonunda amacımız **y** ve **X** verildiğinde

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

eşitliğini sağlayan **β** katsayılarını bulmaktır. Bunu yapmak için hata kareler toplamını minimize ederiz. **X** full rank ise tek çözüm vardır.

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_{\mathbf{2}}^{\mathbf{2}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_{\mathbf{2}}^{\mathbf{2}}$$

$\|.\|_2$ Euclid normu (L_2norm)

OLS Regression

Başka bir ifadeyle p değişken ve n gözlem olduğunda

$$\boldsymbol{RSS} = \sum_{i=1}^{n} (\boldsymbol{y_i} - \widehat{\boldsymbol{y_i}})^2 = \sum_{i=1}^{n} \left(\boldsymbol{y_i} - \boldsymbol{\beta_0} - \sum_{j=1}^{p} \boldsymbol{x_{ij}} \boldsymbol{\beta_j} \right)^2$$

Tam çoklu bağlantı olduğunda **X** full rank olmadığından tek çözüm yoktur.
p>>n olduğunda yine **X** full rank olmayabilir.

OLS Regression

- OLS tahminleri genellikle yetersizdir.
- OLS tahminlerinde bias azdır fakat varyans fazladır.
- Bazı katsayılar küçültülerek ya da sıfırlanarak tahmin geçerliliği (prediction accuracy) arttırılabilir.
- Alt küme seçimi yapmak da bir yöntemdir. Fakat bu yöntemde sabit modeller elde etmek zordur.

Ridge Regression- L_2 Penalty - 1970

- OLS deki parametre kestirimleri yansızdı. Bazen kestirimlerin yanlı olması daha iyi tahmin değerleri elde etmeyi sağlar.
- OLS üzerine bir L_2 penalty konmasıyla Ridge regression elde edilir.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \|\boldsymbol{\lambda\beta}\|_2^2 \quad \boldsymbol{ya\ da}$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 \ \boldsymbol{s.t.}\ \|\boldsymbol{\lambda\beta}\|_2^2 < \boldsymbol{t}$$

Ridge Regression

- Ridge tahminleri daha iyidir.
- Çoklu bağlantı sorunu çözülür.
- Katsayılar küçülmüştür. (Shrinkage)
- MSE, OLS dekine göre daha küçük olabilir.
- Katsayılar değişkenlerin birimlerine duyarlı olduğundan standartlaştırma yapılır.

Fakat

- Tüm değişkenler modelde kalmıştır.

The Lasso - L_1 Penalty - 1996

- Least Absolute Shrinkage and Selection Operator
- Shrinkage + Subset Selection

$$\hat{\boldsymbol{\beta}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|_{\mathbf{2}}^{\mathbf{2}} + \|\boldsymbol{\lambda\beta}\|_{\mathbf{1}})$$

$\|.\|_1$ Mutlak değer (L_1 norm)

Lasso

- Bazı katsayılar sıfırlanmıştır. (Sparse solution)
- L_1 normu olduğundan türevlenebilirlik kaybolmuştur ve sonucun açık formülü yoktur.
- Bu yüzden algoritmalar geliştirilmiştir.

1996 Quadratic Programming

2003 LARS

2008 Coordinate Descent

Lasso

- p>n durumunda Lasso en çok n tane değişken seçer.
- İkili korelasyonu yüksek bir grup değişken içerisinden sadece birini alır, hangisini seçtiğini önemsemez.
- n>p olduğunda değişkenler arasında yüksek korelasyon varsa Ridge regresyonun Lassoya göre daha iyi tahmin performansı olduğu gözlenmiştir.

Gen Seçimi

- Gen seçimi durumunda p>>n dir.
- Bazı genlerin yüksek korelasyonlu gruplar halinde aynı biyolojik yolu paylaştığı düşünülür.
- İdeal gen seçiminde önemsiz genler silinmeli, gruptan bir gen seçilmişse diğerleri de modele girmelidir.
- Lasso bu durum için ideal değildir.

Bridge Regression - L_q Penalty 1993

- Ridge ve Lassonun genelleştirilmiş hali

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|_{\mathbf{2}}^{\mathbf{2}} + \|\boldsymbol{\lambda\beta}\|_{\boldsymbol{q}}^{\boldsymbol{q}})$$

$\|.\|_q$ L_q norm

- 1<q<2 için Elastic Nete benzer.
- Ridge de olduğu gibi değişken seçimi yapmaz.

Elastic Net - L_1&L_2 Penalty - 2004

- L_1 kısmı değişken seçimi yapar
- L_2 kısmı

(i) Seçilen değişkenlerdeki n adet sınırını kaldırır

(ii) Gruplama etkisini güçlendirir

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|_{\mathbf{2}}^{\mathbf{2}} + \|\boldsymbol{\lambda_1\beta}\|_{\mathbf{1}} + \|\boldsymbol{\lambda_2\beta}\|_{\mathbf{2}}^{\mathbf{2}})$$

Naive Elastic Net

$$\boldsymbol{\alpha} = \frac{\lambda_2}{(\lambda_1+\lambda_2)} \quad \text{seçilirse}$$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{arg}\min_{\boldsymbol{\beta}}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|_{\mathbf{2}}^{\mathbf{2}}) \;\; \boldsymbol{s.t.}\,(\mathbf{1} - \boldsymbol{\alpha})\|\boldsymbol{\beta}\|_{\mathbf{1}} + \boldsymbol{\alpha}\|\boldsymbol{\beta}\|_{\mathbf{2}}^{\mathbf{2}} \le \boldsymbol{t}$$

- α=1 durumunda Ridge regression, α=0 olma durumunda Lasso olur.
- Elastic Net, Ridge ve Lassonun konveks bir birleşimidir.

Elastic Net

- Naive Elastic Net bir Lasso problemi olarak yazılabilir.

$$\widehat{\boldsymbol{\beta}}^* = \boldsymbol{arg}\min_{\boldsymbol{\beta}^*}(\|\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_{\mathbf{2}}^{\mathbf{2}} + \frac{\boldsymbol{\lambda_1}}{\sqrt{\mathbf{1} + \boldsymbol{\lambda_2}}}\|\boldsymbol{\beta}^*\|_{\mathbf{1}})$$

- Elastic Net katsayıları ile Naive Elastic Net katsayıları arasındaki ilişki aşağıdaki gibidir.

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{enet}) = (\mathbf{1} + \boldsymbol{\lambda_2})\widehat{\boldsymbol{\beta}}\ (\boldsymbol{naive\ enet})$$

- Elastic Net, Naive Elastic Nete göre daha performanslıdır.

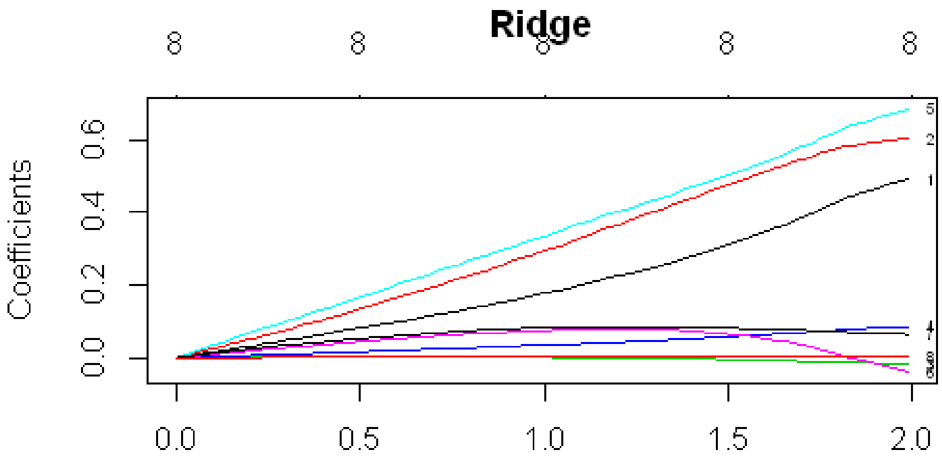
Elastic Net

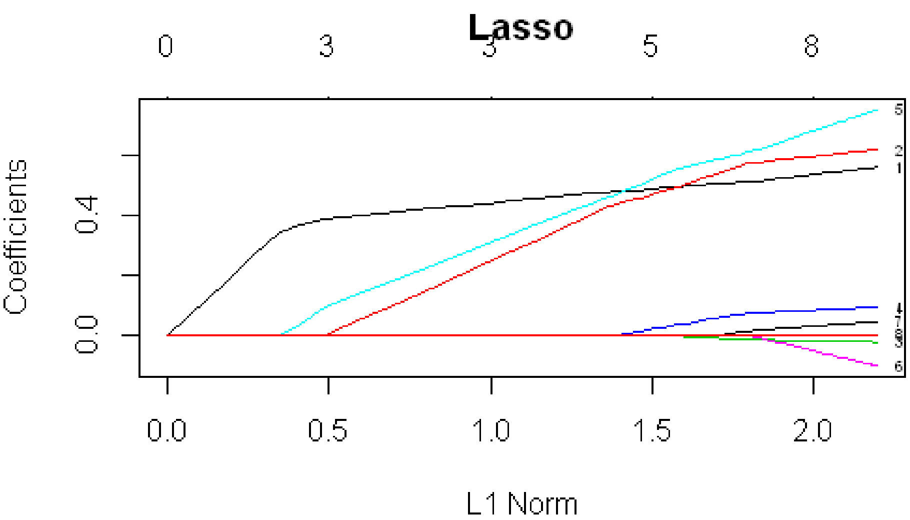
- Elastic Net iyi tahmin performansına sahip, grup etkisi içeren bir model üretir.
- Elastic Net alt küme seçimi yapar (Sparse solution).
- LARS ve Coordinate descent algoritmaları ile hesaplanabilir.
- Lassodan daha performanslıdır.
- Model için iki parametrenin seçilmesi gereklidir.

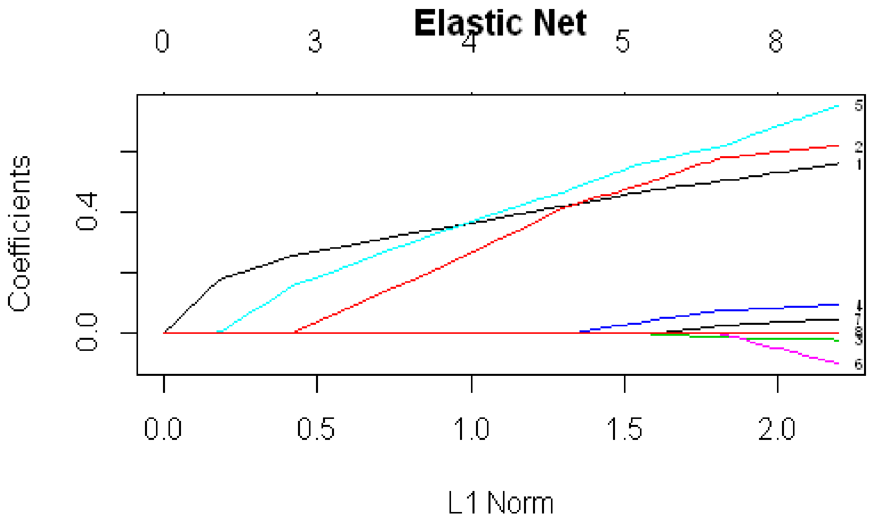
Uygulama

- Prostat kanseri verisi
- Bir bağımlı değişken logpsa ve 8 adet bağımsız değişken
- R glmnet paketi
- Elastic net için α=0.5 seçildi.

Coefficient Paths



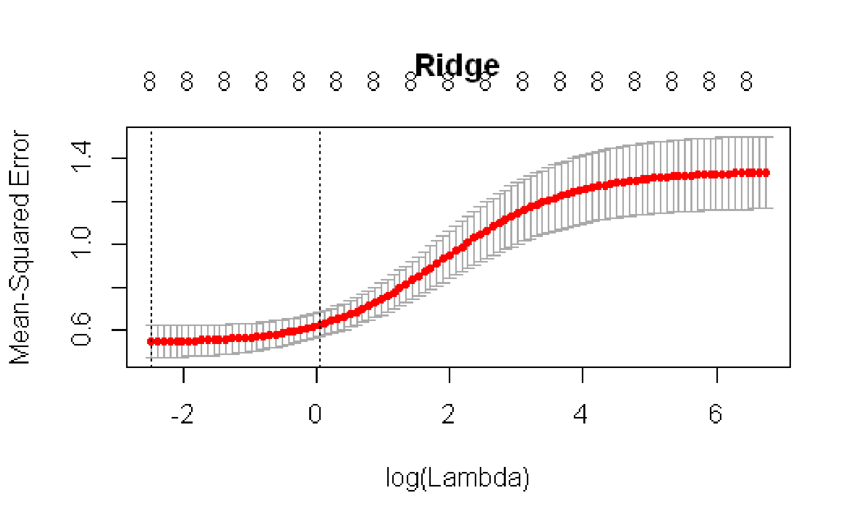




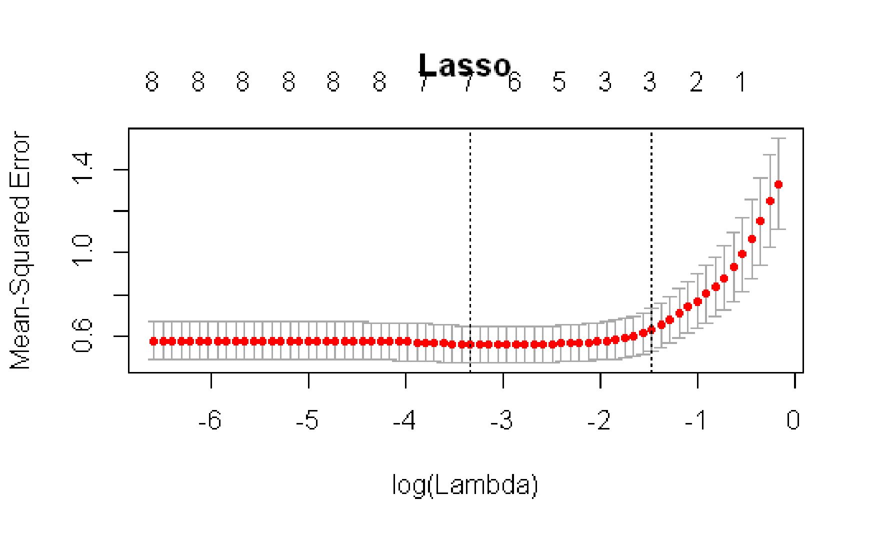
Katsayılar

Ridge	Lasso	ENet
(Intercept) 0.012 lcavol 0.492 lweight 0.604 age -0.017 lbph 0.086 svi 0.685 lcp -0.040 gleason 0.064 pgg45 0.003	(Intercept) 0.154 lcavol 0.507 lweight 0.546 age -0.008 lbph 0.062 svi 0.590 lcp . gleason 0.001 pgg45 0.002	(Intercept) 0.120 lcavol 0.497 lweight 0.566 age -0.011 lbph 0.070 svi 0.608 lcp . gleason 0.021 pgg45 0.002

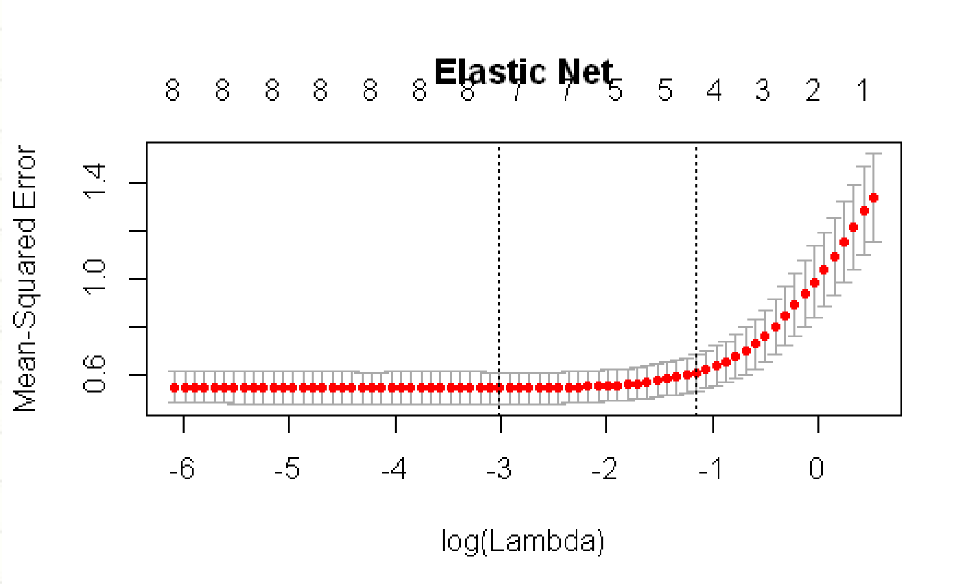
MSE



Ridge
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
Mean-Squared Error
0.6
1.0
1.4
-2
0
2
4
6
log(Lambda)



Lasso
8 8 8 8 8 8 7 7 6 5 3 3 2 1
Mean-Squared Error
0.6
1.0
1.4
-6
-5
-4
-3
-2
-1
0
log(Lambda)



Elastic Net
8 8 8 8 8 8 8 7 7 5 5 4 3 2 1
Mean-Squared Error
0.6
1.0
1.4
-6
-5
-4
-3
-2
-1
0
log(Lambda)