UVA CS 6316 – Fall 2015 Graduate: Machine Learning

Lecture 6: Linear Regression Model with Regularizations

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

Where we are ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

Today ➔ Regression (supervised)

- ❑ Four ways to train / perform optimization for linear regression models
  - ❑ Normal Equation
  - ❑ Gradient Descent (GD)
  - ❑ Stochastic GD
  - ❑ Newton's method

- ❑ Supervised regression models
  - ❑ Linear regression (LR)
  - ❑ LR with non-linear basis functions
  - ❑ Locally weighted LR
  - ❑ LR with Regularizations

Today

- ❑ Linear Regression Model with Regularizations
  - ❑ Ridge Regression
  - ❑ Lasso Regression
  - ❑ Elastic net

Review: Vector norms

A norm of a vector ||x|| is informally a measure of the "length" of the vector.

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

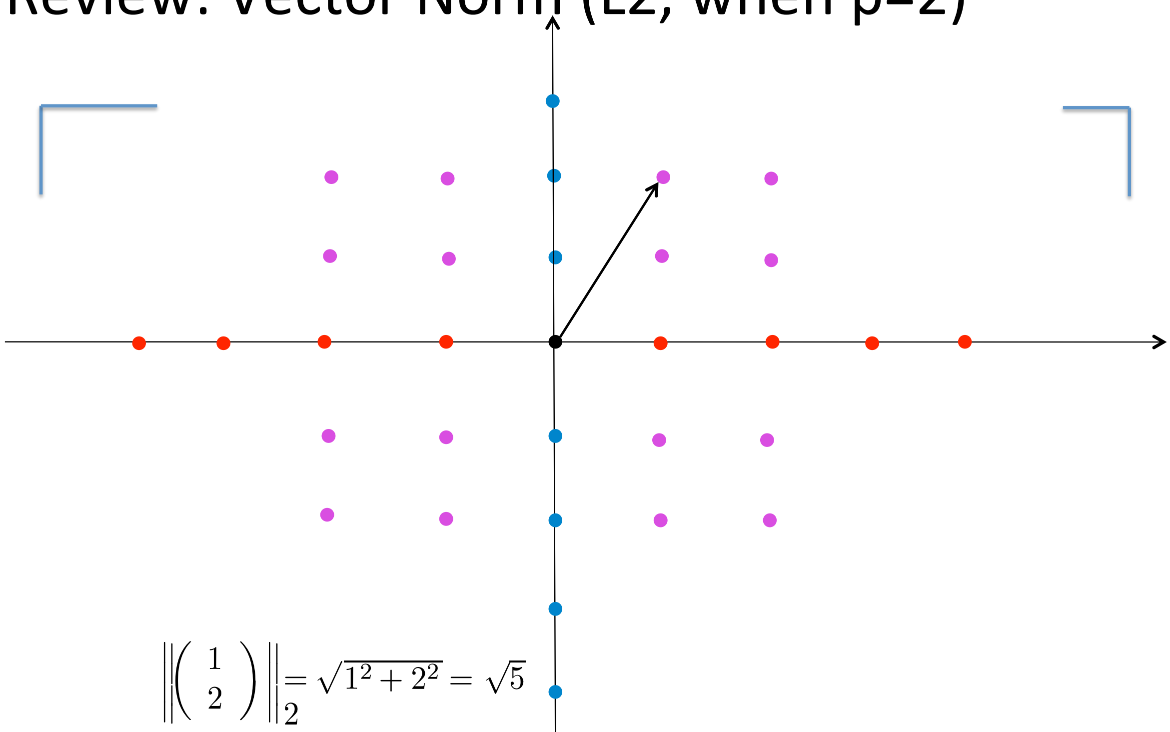
– Common norms: L_1, L_2 (Euclidean)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

– $L_{infinity}$

$$\|x\|_\infty = \max_i |x_i|$$

Review: Vector Norm (L2, when p=2)



$$\left\| \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\|_2 = \sqrt{1^2 + 2^2} = \sqrt{5}$$

Review: Normal equation for LR

- Write the cost function in matrix form:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^T\theta - y_i)^2$$
$$= \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y})$$
$$= \frac{1}{2}\left(\theta^T X^T X\theta - \theta^T X^T \vec{y} - \vec{y}^T X\theta + \vec{y}^T \vec{y}\right)$$

$$\mathbf{X} = \begin{bmatrix} -- & \mathbf{x}_1^T & -- \\ -- & \mathbf{x}_2^T & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_n^T & -- \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

To minimize $J(\theta)$, take derivative and set to zero:

$$\Rightarrow \boxed{X^T X\theta = X^T \vec{y}}$$

The normal equations

$$\Downarrow$$

$$\theta^* = \left(X^T X\right)^{-1} X^T \vec{y}$$

Assume that $X^T X$ is invertible

Comments on the normal equation

$X_{n \times p}$

- In most situations of practical interest, the number of data points N is larger than the dimensionality p of the input space and the matrix $\mathbf{X}$ is of full column rank. If this condition holds, then it is easy to verify that $X^T X$ is necessarily invertible. $p \times p$

$n >> p$ $\quad$ $\text{rank}(X) \leq \min(n, p)$

- The assumption that $X^T X$ is invertible implies that it is positive definite (➔ SSE convex), thus the critical point we have found is a minimum.

- What if $\mathbf{X}$ has less than full column rank? → regularization (later).

Ridge Regression / L2

- If not invertible, a solution is to add a small element to diagonal

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$ Basic Model,

$$\beta^* = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

- The ridge estimator is solution from

HW2

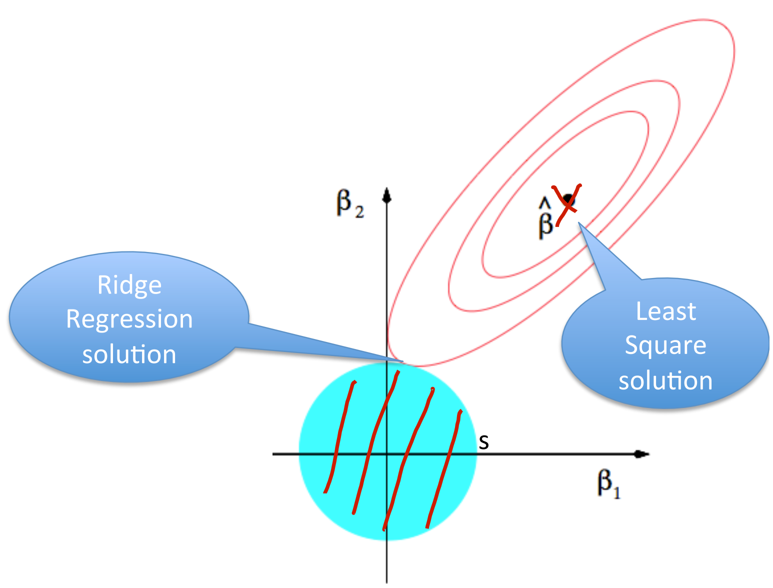
$$\hat{\beta}^{ridge} = \operatorname{argmin}(y - X\beta)^T (y - X\beta) + \lambda\beta^T\beta$$

to minimize, take derivative and set to zero

- Equivalently

$$\hat{\beta}^{ridge} = \arg\min (y - X\beta)^T (y - X\beta)$$
subject to $\sum \beta_j^2 \le s$

Objective Function's Contour lines from Ridge Regression



(1) Ridge Regression / L2

- The parameter $\lambda > 0$ penalizes β_j proportional to its size β_j^2
- Solution is $\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y$
- where I is the identity matrix.
- Note $\lambda = 0$ gives the least squares estimator;
- if $\lambda \rightarrow \infty$, then $\hat{\beta} \rightarrow 0$

Today

- ❑ Linear Regression Model with Regularizations
  - ❑ Ridge Regression
  - ❑ Lasso Regression
  - ❑ Elastic net

(2) Lasso (least absolute shrinkage and selection operator) / L1

- The lasso is a shrinkage method like ridge, but acts in a nonlinear manner on the outcome y.
- The lasso is defined by

$$\hat{\beta}^{lasso} = \arg\min (y - X\beta)^T (y - X\beta)$$

$$\text{subject to } \sum \left|\beta_j\right| \leq s$$

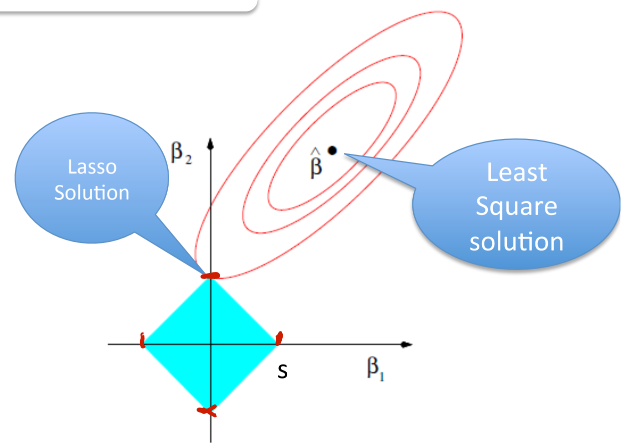
Lasso (least absolute shrinkage and selection operator)

- Notice that ridge penalty $\sum \beta_j^2$ is replaced by $\sum \left|\beta_j\right|$
- Due to the nature of the constraint, if tuning parameter is chosen small enough, then the lasso will set some coefficients exactly to zero.

Lasso (least absolute shrinkage and selection

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}}\left\{\frac{1}{2}\sum_{i=1}^{N}\Big(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

- Suppose in 2 dimension
- $\beta = (\beta_1, \beta_2)$
- $|\beta_1| + |\beta_2| = \text{const}$
- $|\beta_1| + |-\beta_2| = \text{const}$
- $|-\beta_1| + |\beta_2| = \text{const}$
- $|-\beta_1| + |-\beta_2| = \text{const}$



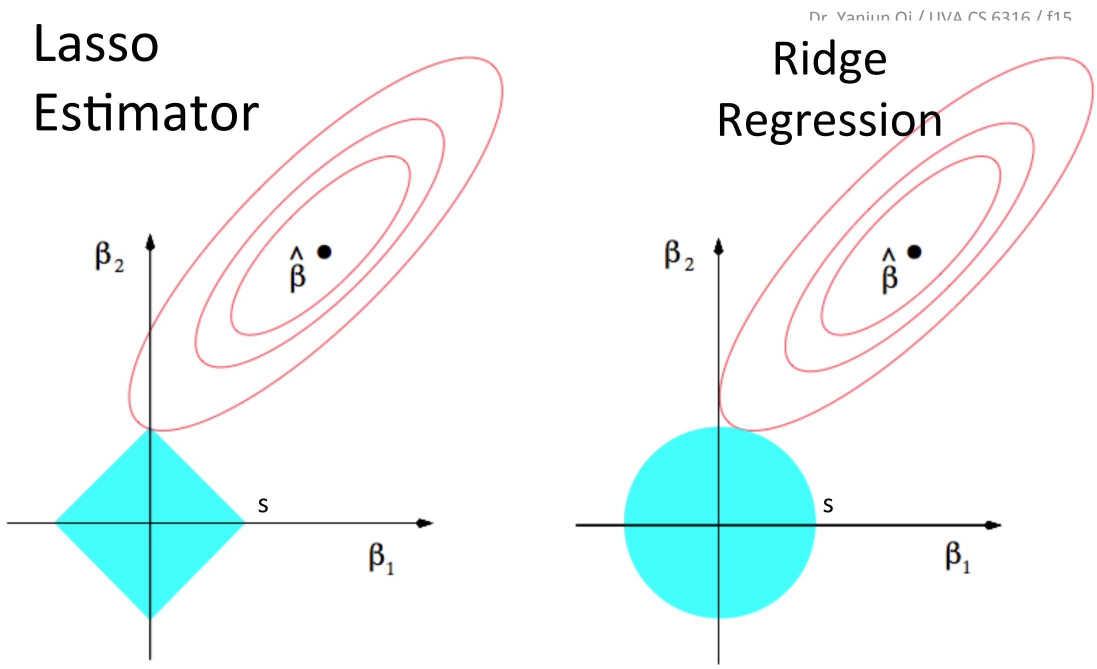


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Today

- ❑ Linear Regression Model with Regularizations
  - ❑ Ridge Regression
  - ❑ Lasso Regression
  - ❑ Elastic net

(3) Hybrid of Ridge and Lasso

Elastic Net regularization

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1$$

many $\beta_j = 0$

- The ℓ_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages *grouping effect*;
  - Stabilizes the ℓ_1 regularization path.

Movie Reviews and Revenues: An Experiment in Text Regression, Proceedings of HLT '10 Human Language Technologies:

III. Model

- Linear regression with the elastic net (Zou and Hastie, 2005)

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta})}{\operatorname{argmin}} \frac{1}{2n}\left[\sum_{i=1}^{n}\left(y_i - (\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})\right)^2\right] + \lambda P(\boldsymbol{\beta})$$

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p}\left(\tfrac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right)$$

Use linear regression to directly predict the opening weekend gross earnings, denoted y, based on features x extracted from the movie metadata and/or the text of the reviews.

More: A family of shrinkage estimators

$$\beta = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

$$\text{subject to} \quad \sum \left|\beta_j\right|^q \le s$$

- for q >=0, contours of constant value of $\sum_j \left|\beta_j\right|^q$ are shown for the case of two inputs.

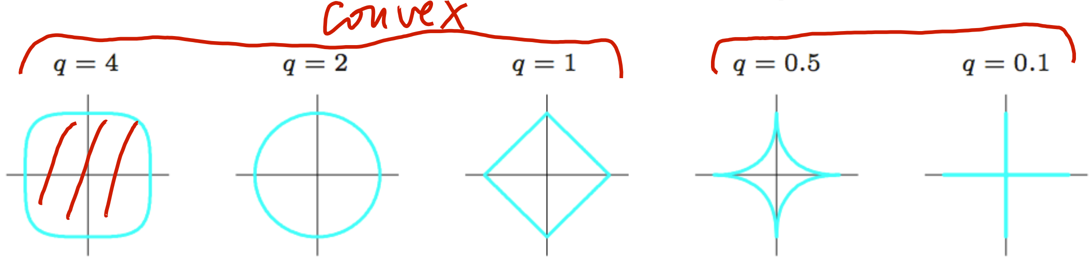
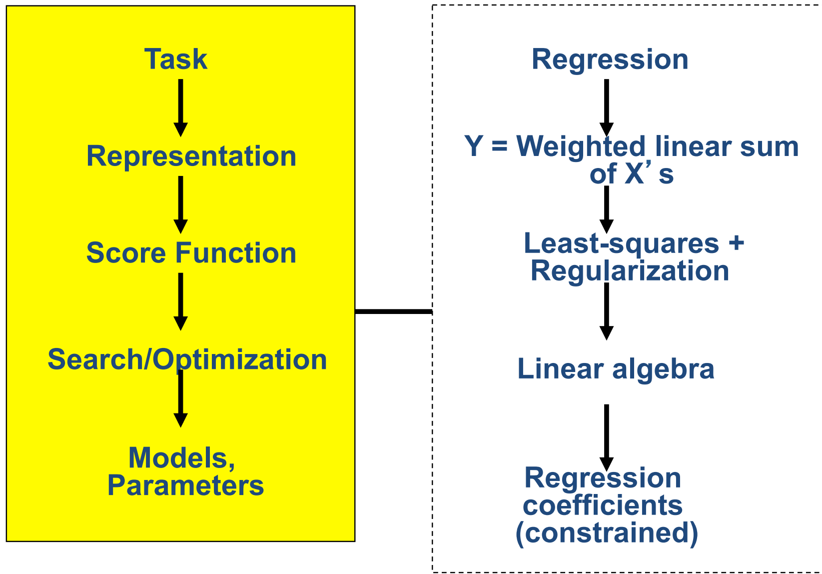


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Regularized multivariate linear regression



Summary: Regularized multivariate linear regression

- Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$
  - LR estimation: $\min SSE = \sum \left(Y - \hat{Y} \right)^2$
  - LASSO estimation: $\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \sum_{j=1}^{p} |\beta_j|$
  - Ridge regression estimation: $\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \sum_{j=1}^{p} \beta_j^2$

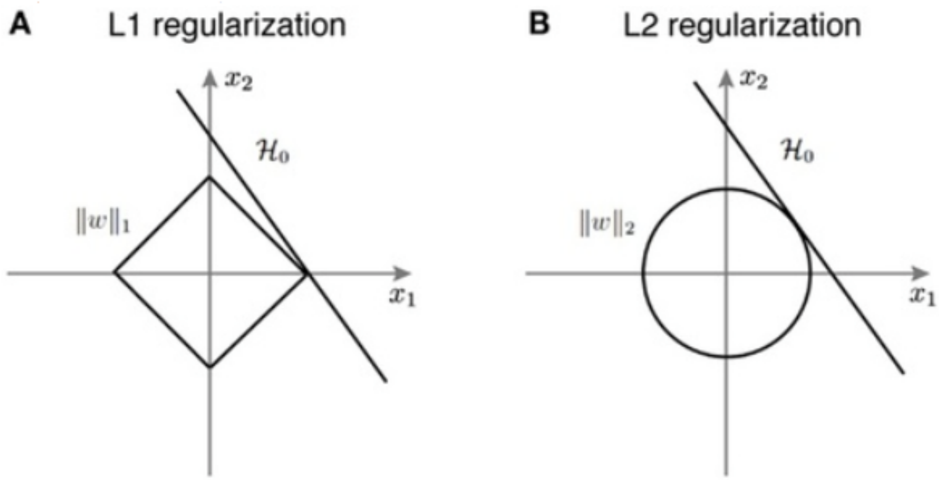
22/54

Error on data + Regularization

EXTRA (NOT REQUIRED)

Today

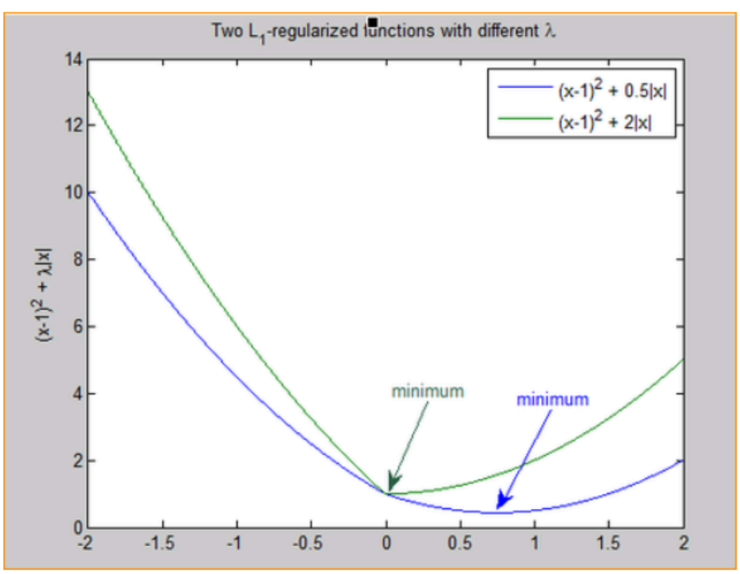
- ❑Linear Regression Model with Regularizations
  - ❑ Ridge Regression
  - ❑ Lasso Regression
    - ❑ Extra: how to perform training
  - ❑ Elastic net



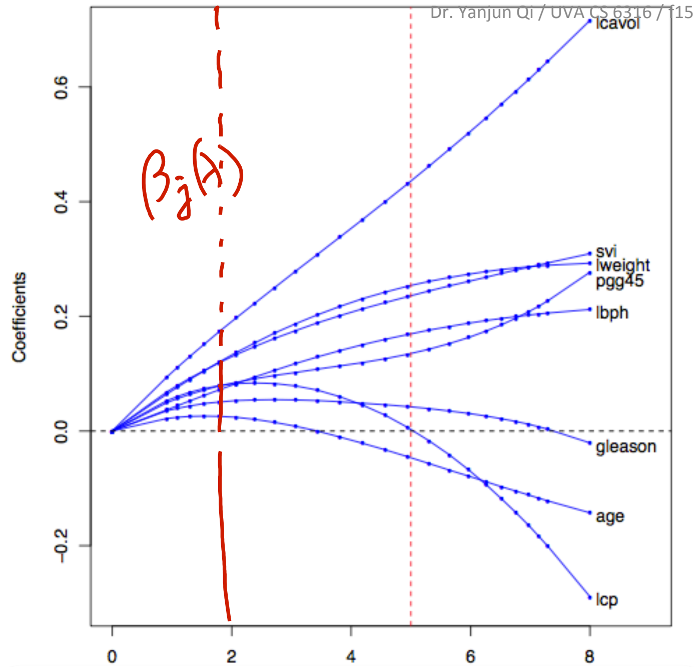
due to the nature of L_1 norm, the viable solutions are limited to the corners, which are on one axis only - in the above case x1. Value of x2 = 0. This means that the solution has eliminated the role of x2 leading to sparsity

L_1-regularized loss function $F(x) = f(x) + \lambda\|x\|_1$ is non-smooth. It's not differentiable at 0. Optimization theory says that the optimum of a function is either the point with 0-derivative or one of the irregularities (corners, kinks, etc.). So, it's possible that the optimal point of F is 0 even if 0 isn't the stationary point of f. In fact, it would be 0 if λ is large enough (stronger regularization effect). Below is a graphical illustration.

http://www.quora.com/What-is-the-difference-between-L1-and-L2-regularization

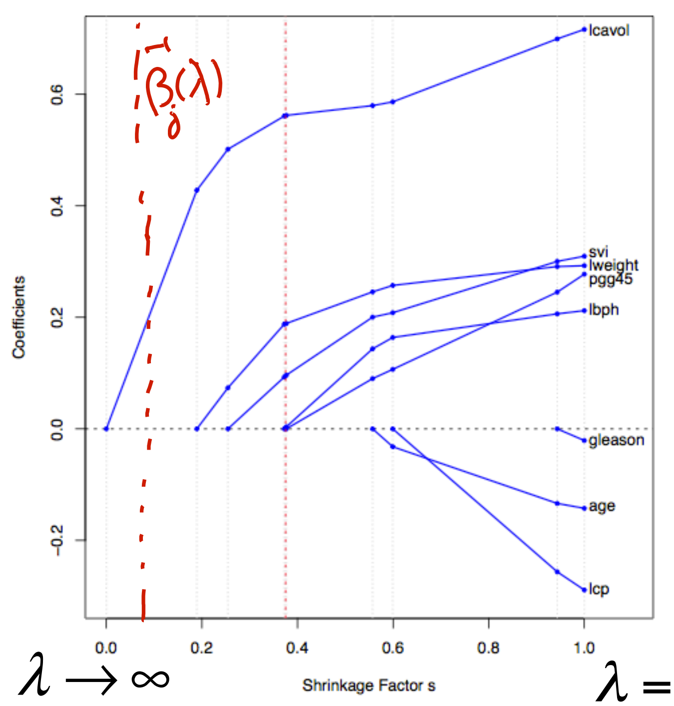


Regularization path of a Ridge Regression



$\lambda \to \infty$ $\lambda = 0$

Regularization path of a Lasso Estimator



$\lambda \to \infty$ $\lambda = 0$

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

How to Learn Parameter for Lasso

$$\hat{\beta}^{lasso} = \arg\min (y - X\beta)^T (y - X\beta)$$
$$\text{subject to } \sum |\beta_j| \leq s$$

- **ℓ_1-norm is non differentiable!**
  - cannot compute the gradient of the absolute value

$\Rightarrow$ **Directional derivatives** (or subgradient)

$$\text{RSS-loss}(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$
$$= \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$
$$= \left[\sum_{i=1}^{n} (y_i - \underbrace{x_{ij}\beta_j} - x_{i\{-j\}}^T \beta_{-j})^2 \right] + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}$$

if $\beta = (\beta_1, \beta_2, \beta_3)$

$\Rightarrow \beta_{-2} = (\beta_1, \beta_3)$

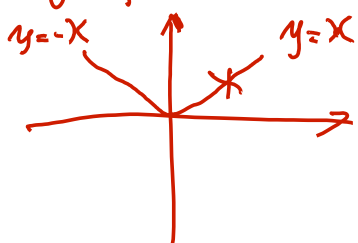
$$\Rightarrow \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \underbrace{2(y_i - x_{ij}\beta_j - x_{i\{-j\}}^T \beta_{-j})} \underbrace{(-x_{ij})}$$
$$+ \lambda \frac{\partial}{\partial \beta_j} |\beta_j|$$

$$= \underbrace{2\sum_{i=1}^{n} x_{ij}^2}_{a_j} \beta_j \underbrace{- 2\sum_{i=1}^{n} \left(y_i - x_{i,[-j]}^T \beta_{[-j]}\right) x_{ij}}_{c_j} + \lambda \frac{\partial}{\partial \beta_j} |\beta_j|$$

$$= a_j \beta_j - c_j + \lambda \frac{\partial}{\partial \beta_j} |\beta_j| \ \underline{\underline{\text{set to}}}\ 0$$

convext $\Rightarrow$ unique

⊛ $y = f(x) = |x|$



$$\partial f(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$$

$$\frac{\partial l}{\partial \beta_j} = \begin{cases} a_j \beta_j - c_j - \lambda \overset{\text{set}}{=} 0, & \text{if } \beta_j < 0 \\ a_j \beta_j - c_j + \lambda, & \underline{\text{if } \beta_j > 0} \\ [a_j \beta_j - c_j - \lambda,\ a_j \beta_j - c_j + \lambda], & \text{if } \beta_j = 0 \end{cases}$$

Set to 0

soft thresholding

$$\hat{\beta}_j = \begin{cases} \dfrac{c_j + \lambda}{a_j}, & \text{if } c_j + \lambda < 0 \Rightarrow c_j < -\lambda \\ \dfrac{c_j - \lambda}{a_j}, & \text{if } c_j > \lambda \\ 0, & \text{if } -\lambda \le c_j \le \lambda \end{cases}$$

Coordinate descent based Learning of Lasso

1. Initialize β
2. Repeat until converged
3. For $j = 1, 2, \ldots, P$ do

$$a_j = 2 \sum_{i=1}^{n} x_{ij}^2$$

$$c_j = 2 \sum_{i=1}^{n} x_{ij} (y_i - x_i^T \beta + x_{ij} \beta_j)$$

if $c_j < -\lambda$

$$\beta_j = (c_j + \lambda) / a_j$$

else if, $c_j > \lambda$

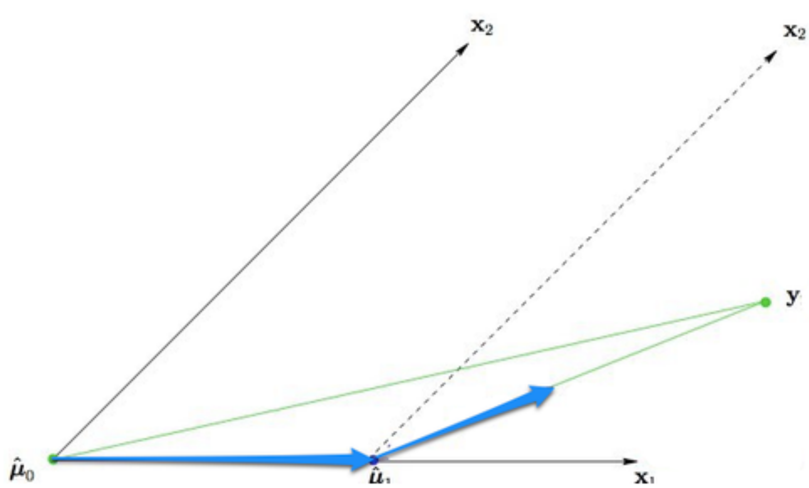
$$\beta_j = (c_j - \lambda) / a_j$$

else, soft-thresholding

$$\beta_j = 0$$

Coordinate descent (WIKI)➔ one does line search along one coordinate direction at the current point in each iteration. One uses different coordinate directions cyclically throughout the procedure.

LARS: Least Angle Regression (state-of-art LASSO solver algorithm)



Lasso when p>n

- Prediction accuracy and model interpretation are two important aspects of regression models.

- LASSO does shrinkage and variable selection simultaneously for better prediction and model interpretation.

Disadvantage:

-In p>n case, lasso selects at most n variable before it saturates

-If there is a group of variables among which the pairwise correlations are very high, then lasso select one from the group

Today

- ❑ Linear Regression Model with Regularizations
  - ❑ Ridge Regression
  - ❑ Lasso Regression
    - ❑ Extra: how to perform training
  - ❑ Elastic net
    - ❑ Extra: how to perform training

(3) Elastic Net: Hybrid of Ridge and Lasso

Elastic Net regularization

$$\hat{\beta} \;=\; \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

- The ℓ_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages *grouping effect*;
  - Stabilizes the ℓ_1 regularization path.

Naïve elastic net

- For any non negative fixed λ_1 and λ_2, naive elastic net criterion:

$$L(\lambda_1, \lambda_2, \beta) = \underbrace{|\mathbf{y} - \mathbf{X}\beta|^2}_{\text{data error}} + \underbrace{\lambda_2|\beta|^2}_{L2} + \underbrace{\lambda_1|\beta|_1}_{L1},$$

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2, \qquad |\beta|_1 = \sum_{j=1}^{p} |\beta_j|.$$

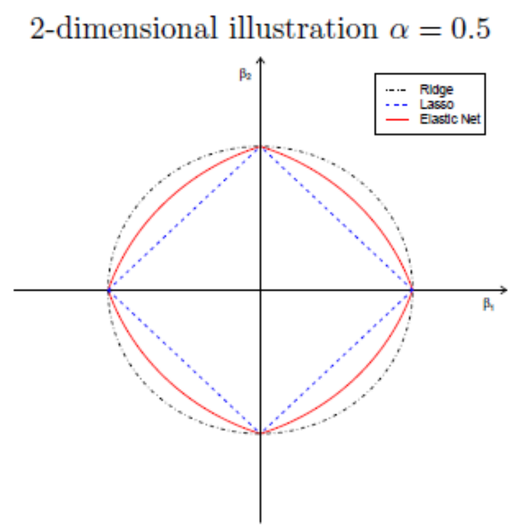
- The naive elastic net estimator is the minimizer of equation

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\}.$$

- Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{y} - \mathbf{X}\beta|^2, \qquad \text{subject to } (1 - \alpha)\,|\beta|_1 + \alpha\,|\beta|^2 \leqslant \boxed{t} \text{ for some } t.$$

Geometry of elastic net



Connecting LASSO and Elastic net

- Lemma: Given (λ_1,λ_2), define an artificial data set (y^*,X^*)

$$\underbrace{\mathbf{X}^*_{(n+p)\times p}}_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\begin{pmatrix} \overset{n\times p}{\mathbf{X}} \\ \sqrt{\lambda_2}\mathbf{I} \end{pmatrix}, \qquad \underbrace{\mathbf{y}^*_{(n+p)}}_{(n+p)\times 1} = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}. \quad n\times 1$$

Let $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ and $\beta^* = \sqrt{(1+\lambda_2)}\beta$. Then the naïve elastic net criterion can be written as

$$L(\gamma, \beta) = L(\gamma, \beta^*) = \left[|\mathbf{y}^* - \mathbf{X}^*\beta^*|^2 + \gamma\,|\beta^*|_1.\right] \Rightarrow \beta^*$$

- Let,

$$\hat\beta^* = \arg\min_{\beta^*} L\{(\gamma, \beta^*)\};$$

- Then

$$\text{elastic } \underbrace{\hat\beta}\, = \frac{1}{\sqrt{(1+\lambda_2)}}\hat\beta^*. \quad \text{lasso augumented}$$

$$X\beta = \vec{Y}$$

$n\times p$ $p\times 1$ $n\times 1$

$(n+p)\times p$ $p\times 1$ $(n+p)\times 1$

Advantage of Elastic net

$p \gg n$

- Native Elastic set can be converted to lasso with augmented data $\Rightarrow X_{n \times p}$
- In the augmented formulation, $\Rightarrow X^*_{(n+p) \times p}$
  - sample size n+p and X^* has rank p
  - ➔ can potentially select all the predictors
- Naïve elastic net can perform automatic variable selection like lasso

Grouping Effect of Naïve Elastic net

Theorem 1. Given data $(\mathbf{y}, \mathbf{X})$ and parameters (λ_1, λ_2), the response $\mathbf{y}$ is centred and the predictors $\mathbf{X}$ are standardized. Let $\hat{\beta}(\lambda_1, \lambda_2)$ be the naïve elastic net estimate. Suppose that $\hat{\beta}_i(\lambda_1, \lambda_2)\,\hat{\beta}_j(\lambda_1, \lambda_2) > 0$. Define

$$D_{\lambda_1,\lambda_2}(i, j) = \frac{1}{|\mathbf{y}|_1} |\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|;$$

then

$$D_{\lambda_1,\lambda_2}(i, j) \leqslant \frac{1}{\lambda_2} \sqrt{\{2(1-\rho)\}},$$

where $\rho = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j$, the sample correlation.

- D is the difference between the coefficient paths of predictors i and j.
- If x_i and x_j are high correlated $\rho=1$, this theorem provides a quantitative description for the grouping effect of Naive Elastic Net.

Elastic Net: Re-scaling of Naive Elastic Net

- **Deficiency of the Naive Elastic Net:** Empirical evidence shows the Naive Elastic Net does not perform satisfactorily. The reason is that there are two shrinkage procedures (Ridge and LASSO) in it. Double shrinkage introduces unnecessary bias.
- Re-scaling of Naive Elastic Net gives better performance, yielding the Elastic Net solution:

$$\hat{\beta}(\text{ENet}) = (1 + \lambda_2) \cdot \hat{\beta}(\text{Naive ENet})$$

- Reason: Undo shrinkage.

Elastic Net: Re-scaling of Naive Elastic Net

Theorem 2. Given data $(\mathbf{y}, \mathbf{X})$ and (λ_1, λ_2), then the elastic net estimates $\hat{\beta}$ are given by

$$\hat{\beta} = \arg\min_{\beta} \beta^{\mathrm{T}} \left(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2} \right) \beta - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\beta + \lambda_1 |\beta|_1. \tag{14}$$

It is easy to see that

$$\hat{\beta}(\text{lasso}) = \arg\min_{\beta} \beta^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X}) \beta - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\beta + \lambda_1 |\beta|_1. \tag{15}$$

Hence theorem 2 interprets the elastic net as a stabilized version of the lasso. Note that $\hat{\Sigma} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a sample version of the correlation matrix Σ and

$$\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2} = (1 - \gamma)\hat{\Sigma} + \gamma \mathbf{I}$$

- Rescaling after the elastic net penalization is mathematically equivalent to replacing Σ with its shrunken version in the lasso.

Computation of Elastic Net

- First solve the Naive Elastic Net problem, then rescale it.
- For fixed λ_2, the Naive Elastic Net problem is equivalent to a LASSO problem, with a huge design matrix if p >> n
- LASSO already has an efficient solver called LARS (Least Angle Regression).

Today Recap

- ❑ Linear Regression Model with Regularizations
  - ❑ Ridge Regression
  - ❑ Lasso Regression
    - ❑ Extra: how to perform training
  - ❑ Elastic net
    - ❑ Extra: how to perform training

Extra: Shrinkage Bias Term ?

- If the data is not centered, there exists bias term
  - http://stats.stackexchange.com/questions/86991/reason-for-not-shrinking-the-bias-intercept-term-in-regression

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- We normally assume we centered x and y. If this is true, no need to have bias term, e.g., for lasso,

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \|\beta\|_1$$

References

- ❑ Big thanks to Prof. Eric Xing @ CMU for allowing me to reuse some of his slides
- ❑ Prof. Nando de Freitas's tutorial slide
- ❑ **Regularization and variable selection via the elastic net,** Hui Zou and Trevor Hastie, *Stanford University, USA*