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

9/30/15

Where we are? ➔
Five major sections of this course

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

9/30/15
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
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)
  $$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$
  $$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$
- $L_{\infty}$
  $$\|x\|_{\infty} = \max_i |x_i|$$

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:
  \[ J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (x_i^T \theta - y_i)^2 \]
  \[ = \frac{1}{2} (X\theta - \bar{y})^T (X\theta - \bar{y}) \]
  \[ = \frac{1}{2} (\theta^T X^T X \theta - \theta^T X^T \bar{y} - \bar{y}^T X \theta + \bar{y}^T \bar{y}) \]

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

\[ \Rightarrow \quad X^T X \theta = X^T \bar{y} \]

Assume that \( X^T X \) is invertible

\[ \theta^* = \left( X^T X \right)^{-1} X^T \bar{y} \]

Comments on the normal equation

- 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 \( X \) is of full column rank. If this condition holds, then it is easy to verify that \( X^T X \) is necessarily invertible.

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

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

- What if \( X \) has less than full column rank? \( \Rightarrow \) regularization (later).
Ridge Regression / L2

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

\[ Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \]

Basic Model,

\[ \beta^* = \left( X^T X + \lambda I \right)^{-1} X^T \bar{y} \]

• The ridge estimator is solution from

\[ \hat{\beta}^{ridge} = \arg\min(y - X\beta)^T(y - X\beta) + \lambda \beta^T \beta \]

to minimize, take derivative and set to zero

• Equivalently

\[ \hat{\beta}^{ridge} = \arg\min(y - X\beta)^T(y - X\beta) \]

subject to \[ \sum \beta_j^2 \leq s \]

Objective Function’s Contour lines from Ridge Regression
(1) Ridge Regression / L2

- The parameter $\lambda > 0$ penalizes $\beta_j$ proportional to its size $\beta_j^2$
- 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 $\lambda \to \infty$, then $\hat{\beta} \to 0$

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

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

Lasso (least absolute shrinkage and selection operator)

- Notice that ridge penalty \(\sum \beta_j^2\) is replaced by \(\sum |\beta_j|\)

- 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)

\[
\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\}.
\]

- 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 \(|\beta_1| + |\beta_2| \leq t\) and \(\beta_1^2 + \beta_2^2 \leq t^2\), 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

$$\hat{\beta} = \arg \min_\beta \| y - X \beta \|^2 + \lambda_2 \| \beta \|^2 + \left( \lambda_1 \| \beta \|_1 \right)$$

- The $\ell_1$ 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 $\ell_1$ 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)

\[
\hat{\theta} = \arg\min_{\theta=(\beta_0, \beta)} \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - (\beta_0 + x_i^T \beta) \right)^2 + \lambda P(\beta)
\]

\[
P(\beta) = \sum_{j=1}^{p} \left( \frac{1}{2} (1 - \alpha) \beta_j^2 + \alpha |\beta_j| \right)
\]

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

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

subject to \( \sum |\beta_j|^q \leq s \)

• for \( q \geq 0 \), contours of constant value of \( \sum_j |\beta_j|^q \) are shown for the case of two inputs.

![Contours](image)

**FIGURE 3.12.** Contours of constant value of \( \sum_j |\beta_j|^q \) for given values of \( q \).
Regularized multivariate linear regression

**Task**
- Representation
- Score Function
- Search/Optimization
- Models, Parameters

**Regression**
- \( Y = \text{Weighted linear sum of } X \text{'s} \)
- Least-squares + Regularization
- Linear algebra
- Regression coefficients (constrained)

**Summary:**
Regularized multivariate linear regression

- **Model:**
  \[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \]

- **LR estimation:**
  \[ \min SSE = \sum (Y - \hat{Y})^2 \]

- **LASSO estimation:**
  \[ \min SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{j=1}^{p} |\beta_j| \]

- **Ridge regression estimation:**
  \[ \min SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{j=1}^{p} \beta_j^2 \]

**Error on data** + **Regularization**
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 $x_1$. Value of $x_2 = 0$. This means that the solution has eliminated the role of $x_2$ leading to sparsity.

Regularization path of a Ridge Regression

Regularization path of a Lasso Estimator

FIGURE 3.10. Profiles of lasso coefficients, as the tuning parameter \( \lambda \) is varied. Coefficients are plotted versus \( s = t/\sum |\beta_j| \). A vertical line is drawn at \( s = 0.35 \), 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

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

- \( \ell_1 \)-norm is non differentiable!
  - cannot compute the gradient of the absolute value
  \[ \Rightarrow \text{Directional derivatives (or subgradient)} \]

\[ \text{RSS Loss} (\lambda) = \sum_{i=1}^{n} (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \]

- \( \beta = (\beta_1, \beta_2, \beta_3) \)
  \[ \Rightarrow \beta_{-2} = (\beta_1, \beta_3) \]

- \[ \frac{\partial}{\partial \beta_j} = \sum_{i=1}^{n} 2(y_i - \hat{y}_i) \hat{y}_i' \beta_j \]
\[ + \lambda \frac{\partial}{\partial \beta_j} |\beta_j| \]
\[
\begin{align*}
\frac{\partial}{\partial \beta_j} &= 2 \sum_{i=1}^{n} x_{ij} \beta_j - 2 \sum_{i=1}^{n} (y_i - x_{ij} \beta_j) \frac{1}{a_j} - \frac{n \beta_j}{\partial \beta_j} |\beta_j| + \lambda \frac{\partial}{\partial \beta_j} |\beta_j| \\
&= a_j \beta_j - C_j + \lambda \frac{\partial}{\partial \beta_j} |\beta_j| \overset{\text{set to } 0}{=} 0
\end{align*}
\]

\[y = f(x) = |x|\]

\[\begin{align*}
9/30/15 & & \text{31}
\end{align*}\]
Coordinate descent based Learning of Lasso

**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)

1. **Initialize** $\beta$
2. **Repeat** until converged
3. **For** $j = 1, 2, \ldots, p$ **do**
   
   $a_j = \sum_{i=1}^{n} x_{ij}^2$
   
   $e_j = \sum_{i=1}^{n} x_{ij} (y_i - x_i^T \beta + x_i j) $
   
   **if** $e_j < -\lambda$
   
   $\beta_j = (e_j + \lambda) / a_j$
   
   **else if** $e_j > \lambda$
   
   $\beta_j = (e_j - \lambda) / a_j$
   
   **else**
   
   $\beta_j = 0$

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.
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**

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

- The $\ell_1$ 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 $\ell_1$ regularization path.

Naïve elastic net

- For any non negative fixed $\lambda_1$ and $\lambda_2$, naïve elastic net criterion:
  \[
  L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1.
  \]
  \[
  |\beta|^2 = \sum_{j=1}^p \beta_j^2, \quad |\beta|_1 = \sum_{j=1}^p |\beta_j|.
  \]
- The naïve elastic net estimator is the minimizer of equation
  \[
  \hat{\beta} = \arg\min_{\beta} \{ L(\lambda_1, \lambda_2, \beta) \}.
  \]

- Let $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$
  \[
  \hat{\beta} = \arg\min_{\beta} |y - X\beta|^2, \quad \text{subject to } (1 - \alpha) |\beta|_1 + \alpha |\beta|^2 < t \text{ for some } t.
  \]
Geometry of elastic net

2-dimensional illustration $\alpha = 0.5$

Connecting LASSO and Elastic net

- Lemma: Given $(\lambda_1, \lambda_2)$, define an artificial data set $(y^*, X^*)$

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

$$L(\gamma, \beta) = L(\gamma, \beta^*) = \|y^* - X^* \beta^*\|_2^2 + \gamma \|\beta^*\|_1 \Rightarrow \beta^*$$

- Let,

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

- Then

$$\hat{\beta} = \frac{1}{\sqrt{1 + \lambda_2}} \hat{\beta}^*_{\text{lasso augmented}}$$
Advantage of Elastic net

- Native Elastic set can be converted to lasso with augmented data
  \[ \Rightarrow X_{n \times p} \]

- In the augmented formulation,
  - sample size \( n+p \) and \( X^* \) has rank \( p \)
  - \( \Rightarrow X_{(n+p) \times p} \)
  - 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 \((y, X)\) and parameters \((\lambda_1, \lambda_2)\), the response \(y\) is centred and the predictors \(X\) are standardized. Let \(\hat{\beta}(\lambda_1, \lambda_2)\) be the naïve elastic net estimate. Suppose that \(\hat{\beta}_i(\lambda_1, \lambda_2) \neq \hat{\beta}_j(\lambda_1, \lambda_2) > 0\). Define

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

then

\[
D_{\lambda_1, \lambda_2}(i, j) \leq \frac{1}{\lambda_2 \sqrt{2(1 - \rho)}},
\]

where \(\rho = x_i^T x_j\), 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.

---

**Theorem 2.** Given data \((y, X)\) and \((\lambda_1, \lambda_2)\), then the elastic net estimates \(\hat{\beta}\) are given by

\[
\hat{\beta} = \arg\min_\beta \beta^T \left( \frac{X^TX + \lambda_2 I}{1 + \lambda_2} \right) \beta - 2y^T X \beta + \lambda_1 |\beta|_1.
\]

(14)

It is easy to see that

\[
\hat{\beta}(\text{lasso}) = \arg\min_\beta \beta^T (X^TX) \beta - 2y^T X \beta + \lambda_1 |\beta|_1.
\]

(15)

Hence theorem 2 interprets the elastic net as a stabilized version of the lasso. Note that \(\hat{\Sigma} = X^TX\) is a sample version of the correlation matrix \(\Sigma\) and

\[
\frac{X^TX + \lambda_2 I}{1 + \lambda_2} = (1 - \gamma) \hat{\Sigma} + \gamma I
\]

- Rescaling after the elastic net penalization is mathematically equivalent to replacing \(\Sigma\) with its shrunken version in the lasso.
Computation of Elastic Net

- First solve the Naive Elastic Net problem, then rescale it.
- For fixed $\lambda_2$, 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


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

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

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

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