2000

Beta Induced Sparsity Algorithm

Ernest Fokoue

Journal of xxx xxxxx xxxxx 1 (2009) 1-48 Submitted 4/00; Published 10/00

Beta Induced Sparsity Algorithm

Ernest Fokoué ERNEST.FOKOUE@RIT.EDU
Center for Quality and Applied Statistics
Rochester Institute of Technology
Rochester, NY 14623, USA

Editor: xxxxxxx

Abstract

We propose a novel technique that exploits some interesting properties of the Beta distribution to derive a sparse solution to the traditional general linear regression under the Gaussian noise assumption. Our proposed technique provides a theoretically, conceptually and computationally better alternative to both the LASSO and the relevance vector machine in the sense that it is centered around an objective function that is convex and easy to interpret. We demonstrate the strength of our proposed technique through examples, and we also provide a theoretical proof of the merits of our method.

Keywords: Normal Linear model, Maximum a posteriori, Sparsity, Beta hyperprior.

1. Introduction

Let $\mathbf{x}_i^\top \equiv (\mathrm{x}_{1i}, \mathrm{x}_{2i}, \cdots, \mathrm{x}_{pi})$ denote the p-dimensional vector of characteristics. Consider the p-dimensional vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^\top$ of regression coefficients. Assuming a linear model without intercept, the response or measurement at point $\mathbf{x}_i$ can be written as

$$Y_j = \beta_1 \mathrm{x}_{1j} + \beta_2 \mathrm{x}_{i2} + \cdots + \beta_p \mathrm{x}_{ip} + \epsilon_i, \qquad i = 1, \cdots, n$$

where ϵ_j will be assumed i.i.d $\mathbf{N}(0, \sigma^2)$ throughout this paper. Under this homoscedastic noise model, the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is such that

$$\hat{\boldsymbol{\beta}}_{\mathsf{OLS}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} \quad \text{and} \quad \mathsf{cov}(\hat{\boldsymbol{\beta}}_{\mathsf{OLS}}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

where $\mathbf{X}$ is the $n \times p$ data matrix, and $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_n)^\top$ is the vector of n observed response values. Recall here that

$$\hat{\boldsymbol{\beta}}_{\mathsf{OLS}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} [\mathrm{y}_i - (\beta_1 \mathrm{x}_{1j} + \cdots + \beta_p \mathrm{x}_{ip})]^2 \right\} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\mathrm{y}_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right\}$$

When the data matrix $\mathbf{X}$ is full rank and well conditioned, the above least squares estimator has many desirable properties. When the data matrix $\mathbf{X}$ is not full rank, one solution is ridge regression

$$\hat{\boldsymbol{\beta}}_{\mathsf{RIDGE}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\mathrm{y}_i - \sum_{j=1}^{p} \beta_j \mathrm{x}_{ij} \right)^2 \right\} + \mathrm{w} \sum_{j=1}^{p} \beta_j^2$$

Explicitly, the ridge estimator is given by

$$\hat{\boldsymbol{\beta}}_{\mathsf{RIDGE}} = \left(\mathbf{X}^\top\mathbf{X} + \mathrm{w}\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

The ridge estimator is biased estimator of $\boldsymbol{\beta}$, but has the advantage of a reduced variance. The other severe limitations is that the regularizer w controls the overall extend of shrinkage and does not have the capability of addressing each variable separately. In other words, if w goes to infinity, all the values of $\boldsymbol{\beta}$ shrink towards 0. When the matrix $\mathbf{X}$ is not full rank, it is clearly the case that some of the predictor variables are not relevant, and one would want an estimator of $\boldsymbol{\beta}$ for which the β_j's corresponding to such variables to be zero, while the β_j's for the relevant variables are nonzero. Hence the idea of generalized ridge, for which the objective function is

$$\hat{\boldsymbol{\beta}}_{\mathsf{GR}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{y}_i - \sum_{j=1}^{p}\beta_j\mathrm{x}_{ij}\right)^2 \right\} + \sum_{j=1}^{p}\mathrm{w}_j\beta_j^2$$

A little bit of algebra reveals that

$$\hat{\boldsymbol{\beta}}_{\mathsf{GR}} = \left(\mathbf{X}^\top\mathbf{X} + \mathbf{W}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

where $\mathbf{W} = \mathsf{diag}(\mathrm{w}_1, \cdots, \mathrm{w}_p)$ with each $\mathrm{w}_j \in (0, +\infty)$. The most important estimation question then becomes, how to determine the optimal w_j's? Now, if we reparameterize in such a way that $\kappa \in (0, +\infty)$ and $\mathrm{w}_j \in (0, 1)$, then our new and indeed improved objective function is

$$\hat{\boldsymbol{\beta}}_{\mathsf{GR}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{y}_i - \sum_{j=1}^{p}\beta_j\mathrm{x}_{ij}\right)^2 \right\} + \lambda\sum_{j=1}^{p}\mathrm{w}_j\beta_j^2 + \frac{1}{2}\sum_{j=1}^{p}\log(\mathrm{w}_j) + \frac{1}{2}\sum_{j=1}^{p}\log(1-\mathrm{w}_j)$$

It is straightforward to see that

$$\frac{\partial L}{\partial \mathrm{w}_j} = \frac{1}{2\mathrm{w}_j} - \frac{1}{2(1-\mathrm{w}_j)} + \lambda\beta_j^2$$

We need to solve $\partial L/\partial \mathrm{w}_j = 0$

$$-2\lambda\beta_j^2\mathrm{w}_j^2 + 2(\lambda\beta_j^2 - 1)\mathrm{w}_j + 1 = 0.$$

If $\lambda = 0$, then all the w_j are equal to $1/2$. Now, $\mathrm{w}_j = 1/2$ interpreted as a probability of occurrence of variable x_j expresses noninformativeness, which can also be translated as no sparsity pressure put on the parameter space. For $\lambda > 0$,

$$\Delta = 4(\lambda\beta_j^2 - 1)^2 + 8\lambda\beta_j^2 = 4(1 + \lambda^2\beta_j^4) > 0$$

Therefore, for $\lambda > 0$, the solution yielding the optimal weight w_j is

$$\mathrm{w}_j = \frac{\lambda\beta_j^2 - 1 + \sqrt{(1+\lambda^2\beta_j^4)}}{2\lambda\beta_j^2}$$

Simplifying the above yields,

$$\mathrm{w}_j = \frac{1}{2} - \frac{1}{2\lambda\beta_j^2} + \sqrt{\frac{1}{4\lambda^2\beta_j^4} + \frac{1}{4}}$$

which can be written as

$$\mathrm{w}_j = \frac{1}{2}\left[1 - \frac{1}{\lambda\beta_j^2} + \sqrt{1 + \frac{1}{\lambda^2\beta_j^4}}\right]$$

The scenarios confirm the ability of this technique to deliver sparse solutions. Indeed, as λ gets larger and larger for a fixed set of β_j's, both $1/2\lambda\beta_j^2$ and $1/4\lambda^2\beta_j^4$ tend to 0. As a consequence,

$$\lim_{\lambda\to\infty} \mathrm{w}_j = 0/1$$

For relevance variables, meaning variables with very large values of β_j, it is immediate to see that both $1/2\lambda\beta_j^2$ and $1/4\lambda^2\beta_j^4$ tend to 0. As a consequence,

$$\lim_{\beta_j\to\infty} \mathrm{w}_j = 1,$$

signifying that our weight performs exact variable selection.

2. Application to regression problems

It is very straightforward to implement this technique by running a two step iterative procedure as follows.

Choose a range for the values of λ, i.e., $\lambda \in [\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$

For each fixed λ, set initial weights and run the following two step updating procedure

Step 1: Obtain the new vector of regression coefficients

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{W}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

Step 2: Update the weights w_j, for $j = 1, \cdots, n$,

$$\hat{\mathrm{w}}_j = \frac{1}{2}\left[1 - \frac{1}{\lambda\hat{\beta}_j^2} + \sqrt{1 + \frac{1}{\lambda^2\hat{\beta}_j^4}}\right]$$

Repeat the above for each λ until convergence is achieved.

Note: Remember to show the beta distribution at 1/2 to justify the choice of the second order penalty!

3. Application to regression problems

- Get the examples from Clarke, Fokoue, Zhang first with simple structure and then with noise variables and correlated variables
- Get example from Fokoue (2008)

4. Bibliography

- Cite Fokoue, Goel (Communication in Statistics, Theory and Methods)
- Cite LASSO
- Cite RVM
- Cite Grandvalet
- Cite other variable selection papers

All the above updates are straightforward and will be provided once time allows me to do so!