Model Sparsity and Feature Selection

1 The "Bet on Sparsity" Principle

book: Elements, 16.2.2

book: Statistics for high dimensional data, introduction

$$Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i$$

Roughly speaking, for High-dimensional statistical inference to achieve reasonable accuracy or asymptotic consistency, we need

$$\log(p)(\text{sparsity}(\beta)) << n$$

2 Forward Selection

Forward selection starts with no feature(variable) in the model, and adds features to the model one at a time. At each step, the feature that can contribute most to the model is added. The procedure is repeated until one new feature can improve the model significantly(defined by some statistical test threshold).

For a complete survey of feature selection methods, see [3].

3 Regularized Linear Models

3.1 Regularized Linear Regression

Consider the linear regression model

$$Y = \beta_0 + x^T \beta$$

Suppose we have N data points, and p features. Each feature is standardized to have mean 0 an variance 1. Regularized Linear Regression solves the following problem:

$$\min_{(\beta_0,\beta)\in\mathcal{R}^{p+1}} [\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P(\beta)],$$

where $\sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2$ is the square loss term, $P(\beta)$ is a penalty term, and λ controls the strength of the penalty.

There are three kinds of commonly used penalties in linear regression:

- $P(\beta) = \frac{1}{2}||\beta||_2^2 = \frac{1}{2}\sum_{j=1}^{p}\beta_j^2$ is called the RIDGE-regression penalty.
- $P(\beta) = ||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$ is called the LASSO penalty.
- $P_\alpha(\beta) = (1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1 = \sum_{j=1}^{p}[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|]$ is called the elastic-net penalty. The elastic-net penalty is a compromise between the ridge-regression penalty and the lasso penalty.

Ridge Regression shrinks the size the regression coefficients. In linear regression, if there are two correlated features, there coefficients can be poorly determined and have high variance. One of them can have a very large positive coefficient, and the other correlated feature can have a very large negative coefficient. They cancel each other. By adding the ridge penalty, the problem is alleviated, as it shrinks the coefficients towards 0. In the extreme case of k identical features, they each get small identical coefficients. So ridge penalty encourages features to borrow strength from each other. From a Bayesian point of view, the ridge regression estimation assumes that β_j has a Gaussian distribution with 0 mean as its prior distribution. And the solution to ridge regression is the mean (or mode) of the posterior distribution.

Lasso behaves differently than Ridge. If there are several correlated features, Lasso tends to pick one and ignore the rest. That is, some features will have coefficients exactly 0. So Lasso can be used to perform continuous feature selection. From a Bayesian point of view, the Lasso penalty corresponds to a Laplace prior.

To illustrate the behaviors of Ridge and Lasso, we write them as constrained optimization problems.

Ridge regression can be equivalently formulated as

$$\hat{\beta}^{ridge} = \arg\min_{\beta}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t$$

There is a one-to-one correspondence between λ and t.

Lasso regression can be equivalently formulated as

$$\hat{\beta}^{lasso} = \arg\min_{\beta}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \leq t$$

Similarly, there is a one-to-one correspondence between λ and t.

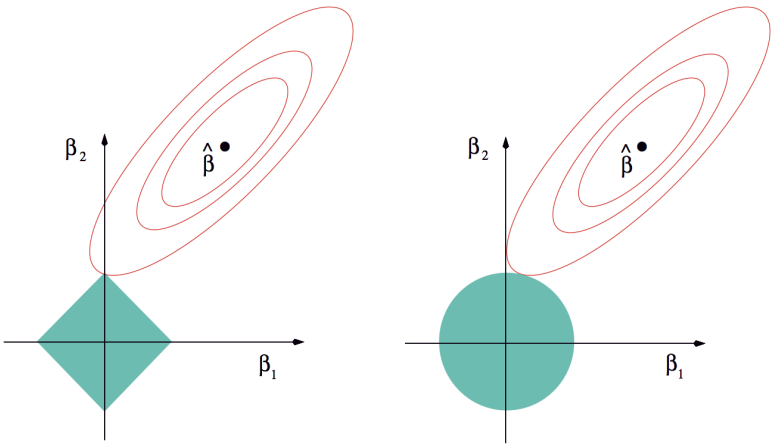

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are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

Figure 1: Source: Figure 3.11 of [4]

Figure 1 shows the difference between lasso and ridge regression estimations when there are only two features. The square loss has elliptical contours. Ridge regression has a disk constraint region, while lasso has a diamond constraint region. In both constrained optimization problems, the optimal solution is the first point where the elliptical contours hit the constraint region. In Lasso regression, if the solution occurs at a corner, then it has one parameter β_j equal to zero. When p is large, there are many corners so that many parameters are likely to become zero.

3.2 Regularized Logistic Regression

The regularized Logistic Regression has the form

$$\max_{\beta_0,\beta}\{\frac{1}{2N}\sum_{i=1}^{N}[y_i(log(P(y=1|x_i)) + (1-y_i)log(P(y=0|x_i)) - \lambda P_\alpha(\beta)\}$$

$$= \max_{\beta_0,\beta}\{\frac{1}{2N}\sum_{i=1}^{N}[y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0+\beta^T x_i})] - \lambda P_\alpha(\beta)\}$$

If we use RIDGE penalty, we get

$$\max_{\beta_0,\beta}\{\frac{1}{2N}\sum_{i=1}^{N}[y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0+\beta^T x_i})] - \lambda\frac{1}{2}\sum_{j=1}^{p}\beta_j^2\}$$

3.3 Solving Regularized Linear Models

packages:

- Liblinear[1]
- glmnet[2]
- sklearn[5]

References

[1] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[2] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[3] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.