MATH 748: Weekly Report

Nov 1-7, 2021

Nayeong Kim (nkim10@sfsu.edu)

Review

The estimator of linear regression is as below.

$$\hat{\beta}^{OLS} = \underset{\beta}{argmin} RSS(\beta) \quad RSS(\beta) = ||Y - X\beta||_2^2$$

Meanwhile, the estimators with penalty terms are as below.

- Lasso: $\hat{\beta}^{LASSO} = \underset{\beta}{argmin} RSS(\beta) + \lambda \sum_j |\beta_j|$
- Ridge: $\hat{\beta}^{LASSO} = \underset{\beta}{argmin} RSS(\beta) + \lambda \sum_j \beta_j^2$

λ is a tuning parameter.

12 Modern Penalized Methods

12.3 Ridge Regression

The solution of Ridge regression is $\hat{\beta}^{Ridge} = (X^T X + \lambda I_p)^{-1} X^T Y$ while the solution of original linear regression is $(X^T X)^{-1} X^T Y$. In the special case of orthonormal design matrix:

$$\hat{\beta}^{Ridge} = \frac{\hat{\beta}^{OLS}}{1+\lambda}$$

We call this "shirnkage". The shirnkage is in different form compared to Lasso regression.

(3) Existence Theorem
Recall $MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = bias^2 + Var$. Assume that the true model is $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$. Then OLS estimator is unbiased. I.e. $E_\beta(\hat{\beta}^{OLS}) = \beta$. On the other hand, both Lasso and Ridge estimators are biased. Hence they reduce the variance and perform better.
There always exists λ such that $MSE(\beta_\lambda^{Ridge} < MSE(\hat{\beta}^{OLS})$.

(4) Bayes interpretation of Lasso and Ridge

- Prior Distribution: $\beta \sim p(\beta)$
- Posterior Distribution: $p(\beta|Y) \propto f(Y|\beta)p(\beta)$
- Posterior Mode: $\hat{\beta}^* = \underset{\beta}{argmax} p(\beta|Y)$
  (This get influence from f)

Let us see the posterior mode is the same as Lasso or Ridge estimator with some constraints.

If ε_i is iid with $N(0,\sigma^2)$ and call the pdf as f. Let $p(\beta) = \prod g(\beta_j)$. Assume that $g \sim N(0,\lambda)$. Then $\hat{\beta}^* = \hat{\beta}_\lambda^{Ridge}$. If $g \sim Laplace(mean = 0, state = \lambda)$, then $\hat{\beta}^* = \hat{\beta}^{Lasso}$.

12.4 Comparison of Lasso and Ridge

Optimizing Lasso estimator is equivalent to minimize $RSS(\beta)$ with the constraint that $\sum|\beta_j| \leq t$ for some t. The shape of the constraint is a square(or a cube) in visualization. t is determined by λ hence different λs show different results.
Optimizing Ridge estimator is equivalent to minimise $RSS(\beta)$ with the constraint $\sum\beta_j^2 \leq t$ for some t. The shape of the constraint is a circle(or a sphere) with the radius $\sqrt{t}$ in visualization.

12.5 Beyond Lasso and Ridge: Other penalties

$\min_\beta RSS + \lambda J(\beta)$ How can we decide a good J?

(1) Fan and Li, 2002 [1]
Good properties are as below.

- "Nearly" unbiased: $E\hat{\beta}) = \beta + o(1)$
- Sparsity
- Robust against small perturbation.

How to make this happen?
(a) Sufficient condition for unbiasedness.
$J'(|\beta|) = 0$ for large $|\beta|$. I.e. penalty is bounded by a constant.
(b) Necessary and sufficient condition for sparsity.
$J(|\beta|)$ is singular at 0.

(2) Adaptive Lasso, Zou, 2006 [2]
$\hat{\beta}^{a.Lasso} = \underset{\beta}{argmin} RSS + \lambda\sum w_j|\beta_j|$
Ideally, small penalty should be applied to large coefficient while large penalty should be applied to small coefficients.
Weights are chosen from the data. Large β_j receive small w_j and small ones receive large.

(3) Elastic Net
$\hat{\beta}^{ENET} = \underset{\beta}{argmin} RSS + \lambda_1 J_1 + \lambda_2 J_2$ where J_1 is Lasso penalty and J_2 is Ridge penalty. Or we can express the penalty term as $\lambda(\alpha J_1 + (1-\alpha)J_2$ for some $\alpha \in [0,1]$. There are two goals. One is to eliminate genes(get a sparse solution, related to L1) and the other is to group selection(related to L2).

(4) Adaptive Elastic Net[Zou and Zhang, 2008]
This network gives weights to the terms in the penalty which has the same approach with (3).

13 Model Assessment and Selection

13.1 Introduction

There are two important goals: i. model assessment, ii. model selection. Model assessment is estimating the performance of different models and choose the best model. Model selection is estimating its prediction error on new data after choosing a final model. Common techniques for these are using validation set, cross-validation, and approaching with analytical estimation criteria(AIC, BIC, C_p, Adjusted R^2, etc).

(1) Test Error and Training Error
Test error is from the response on a new observation. On the other hand, training error is the average error on the training data. Test error is more important to decide whether the given model is good. 0 training error doesn't mean a good model because the model can be over-fitted which (generally) causes a bad performance on a new data set.
For the standard, squared loss is commonly used for regression and $0-1$ loss is commonly used for classification.

(2) Bias and Variance Trade-off
Recall the trade-off between bias and variance. Complex models are flexible but it has a large variance. It can cause over-fitting problem. Simple models are less flexible and highly biased but they have small variance. The optimal model is between them. Test error gives a good metric to choose optimal models.

13.2 Validation Set Approach

In this approach, we randomly divide the available samples into two parts: a training set and a validation set(=hold-out set). After fitting the model on training set, validation set is used for observing performance of the fitted model.
The resulting validation-set error provides an estimate of the test error. Error is assessed by MSE in quantitative case or by 0-1 loss in classification case.
We can tune hyper parameters(e.g. λ for penalty term in Lasso or Ridge regression) with this approach. After fitting the model with different values of hyper parameters, they can be evaluated with the error of the model in validation set.
It is simple and easy to implement. But there are some disadvantages. The validation can vary on different sets. And the actual training set for fitting model is only a subset of the given training set.

13.3 LOOCV(Leave-One-Out Cross Validation)

In this approach, n samples of training set is divided into $n-1$ training set and 1 validation set. This process is repeated n times with n distinct validation sets. This is deterministic process without random factor.
Advantages of this approach are following. First, it tends not to overestimate the test error rate as much as the validation set approach does. Second, the result is not random.
It is not expensive in computation for linear regression. We can write the fitted one time using (n):

$$CV_{(n)} = \frac{1}{n}\sum\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2.$$

For general regression models, $H = X(X^TX)^{-1}X^T$ and $h_i = H_{ii}$ are diagonal elements of H. But, for other methods, this approaches is quite expensive.

13.4 k-fold Cross Validation

k-fold cross validation is a kind of compromise between validation set approach and LOOCV. It divide the data set into k different parts. It uses $k-1$ parts to fit the model and 1 part to evaluate the model. The process is repeated for k different validation sets.
It is not deterministic. Unlike LOOCV, there are some random factors to divide samples into k sets. LOOCV is a special case of k-fold where $k = n$.
Let us see the bias-variance trade-off for k-fold CV. LOOCV has less bias than k-fold CV but it has higher variance. In practice, the bias of k-fold CV is light.
There are some remarks. i. $k = 5$ or $k = 10$ are generally used. ii. In practice, to reduce variance, multiple rounds of CV are performed using different partitions.

13.5 Application of Cross Validation

For example, for regression problems, we can apply CV to all the problems with tuning parameters: shrinkage methods to select λ, best subset selection to select the best model, nonparametric methods like smoothing spline, tree, random forest, boosting, etc. For classification examples, CV can be used to select k in KNN and applied to SVM.

References

[1] Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, Fan and Li, 2002

[2] The Adaptive Lasso and Its Oracle Properties, Zou, 2006

[3] On the adaptive elastic-net with a diverging number of parameters, Zou, Zhang, 2008