MATH 748: Weekly Report

Oct 25-31, 2021

Nayeong Kim (nkim10@sfsu.edu)

Review

There are three types of filter selections: filter, wrapper and embedded. Today, we'll discuss the methods in practice.

11 Feature Selection

11.3 Subset Selection

(1) Best Subset Selection

Fit all the possible models and take the best. The drawback of best subset selection is that it is very time consuming: ${}_pC_1 + \cdots + {}_pC_p = 2^p$. The Solution to the large calculation is selecting stepwisely. There are two ways in selection of variables: forward and backward.

(2) Forward Stepwise Selection

This methods adds one variable in one step. If the metric of optimality is smaller than before, stop the process. Return the model having the least AIC.

It is much more efficient than the best subset selection: $1 + \frac{p(p+1)}{2}$. But it doesn't guarantee that the result is the best model. It is greedy search so the selected variable may not be the globally best choice.

(3) Backward Stepwise Selection

It is also a sequential selection.

Method: i. Starting from p predictors. ii. Delete one predictor by comparing Both sequential methods are greedy approach. They're more efficient than e.g. Best subset vs Backward stepwise selection
$p = 5 \quad X_1, X_2, X_3, X_4$, $p = 4$ $p = 2$ Problem when $p > n$: For the first step, it doesn't work well. It's hard to decide what to remove. (4) How to choose the optimal model. RSS, R^2 are not good metrics to decide the optimal model. because of the different numbers of predictors. Hence there are two different approaches. i. (indirectly) Put a penalty in dimension. (to reduce the bias)
ii. (directly) Use test error. (Validation set approach or cross-validation)

11.4 Metrics to Choose the Optimal Model

(1) Mallow's C_p and AIC

Add some penalty to the dimension using the variance.

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{sigma}^2) \quad d = num(\text{predictors})$$

AIC is defined by maximum likelihood.

$$\text{AIC} = -2log\mathscr{L} + 2d \quad \mathscr{L} \text{ is the maximum likelihood function.}$$

In the case of the linear with Gaussian errors, maximum likelihood and least squares are the same so C_p and AIC are equivalent.

(2) BIC

$$\text{BIC} = \frac{1}{n}(\text{RSS} + log(n)d\hat{sigma}^2)$$

In the equation, n is the sample size. It is similar to C_p: giving penalty to the dimension but heavier. Therefore it results in smaller model than C_p.

(3) Adjusted R^2

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

$$\text{where RSS} = \sum(Y_i - \bar{Y})^2, \quad \text{TSS} = \sum(Y_i - \hat{Y}_i)^2$$

After this adjustment, R^2 will not always increase with more predictors in the model. Unlike metrics in (1), (2), larger adjusted R^2 indicates the better models. Denominator part(TSS) doesn't get the impact of the size of the model, and only the numerator part(RSS) get the influence. It doesn't always increase unlike before adjustment. Hence we can use this for subset selections.

(4) Validations and Cross-Validation

Validations can be done by making a separate validation set or grouping training sets into small parts and doing the validation(cross-validation).

It provides a direct estimate of the test error unlike AIC, BIC, C_p and adjusted R^2. It doesn't require an estimate of the error variance $\sigma^2 : \hat{\sigma}^2$. It has less assumptions about the true underlining model.

(5) One-standard-error Rule

First, calculate the standard error of the estimated test MSE for each model size. And select the simplest model for which estimated test error is within one standard error of the lowest point of the curve. If multiple models are equally good, choose the simplest one which is easier to interpret.

R Labs

(1) How to Call Methods of Subset Selection

```
# Best Subset Selection
regfit.full <- regsubsets(Salary~.,Hitters)
reg.summary <- summary(regfit.full)
# Forward
regfit.fwd <-
regsubsets(Salary~.,data=Hitters,nvmax=19,method="forward")
# Backward
regfit.bwd <- regsubsets(Salary~.,data=Hitters,nvmax=19,method="backward")
```

reg.summary contains the result with various metrics: e.g. cp, rss, adjr2, bic, etc.

(2) How to Plot Result

```
names(reg.summary)
reg.summary$rsq
par(mfrow=c(2,2))
plot(reg.summary$rss,xlab="Number of Variables",ylab="RSS",type="l")
plot(reg.summary$adjr2,xlab="Number of Variables",ylab="Adjusted RSq",type="l")
which.max(reg.summary$adjr2)
points(11,reg.summary$adjr2[11], col="red",cex=2,pch=20)
plot(reg.summary$cp,xlab="Number of Variables",ylab="Cp",type='l')
which.min(reg.summary$cp)
points(10,reg.summary$cp[10],col="red",cex=2,pch=20)
which.min(reg.summary$bic)
plot(reg.summary$bic,xlab="Number of Variables",ylab="BIC",type='l')
points(6,reg.summary$bic[6],col="red",cex=2,pch=20)
```

12 Modern Penalized Methods

12.1 Motivation

Let us consider the bias-variance trade-off. It is optimal to minimize $bias^2 + variance$. (We can think of the U-shaped graph.) Then, how can we achieve it?

(1) Idea

In regression, the loss function is $L = RSS = \sum(Y_i - \beta_0 - \beta^t X_i)^2$. So $\hat{\beta}^{OLS} = \underset{\beta}{argmin} RSS$. In logistic regression, we can use log likelihood as loss function.

The idea is adding a penalized term for β in the loss function: $L(\beta; Y, X) + \lambda J(\beta)$. We call λ as a tuning parameter. Then how can this term improve the performance of the model? Next, we will focus on regression setting. (Classification setting is similar.)

(2) The goal of the penalty term is to reduce the variance by reducing the coefficient terms: β_js.

12.2 LASSO Regression

LASSO(Least Absolute Shrinkage and Selection Operator) is one of the general penalized methods.

(1) Penalty term: $J(\beta) = \sum_{j=1}^{p} |\beta_j|$.

Where $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, we want to reduce the variance by reducing the coefficients. We can write

$$\hat{\beta}_\lambda^{LASSO} = argmin(RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|) \tag{1}$$

(2) Remarks

- Two folds of objectives: i. small RSS ii. small $\sum |\beta_j| = ||\beta||$
- λ controls the relative impact of two terms.
- e.g. $\lambda = 0 \Rightarrow \hat{\beta}_\lambda^{LASSO} = \hat{\beta}^{OLS}$
- e.g. $\lambda = \infty \Rightarrow \beta = 0$
- When λ is large, the penalty term dominates so we will have more shrinkage.
- When λ is small, the original loss term dominates so we will have less shrinkage.
- LASSO yields to *sparse* model.

Why LASSO penalty force coefficients to be zero? In the textbook Chapter 6.2, it talks about the reason. Optimizing (1) is equivalent to the following: $\underset{\beta}{minRSS}$ with constraint $\sum |\beta_j| \leq s$.

(3) How to select λ

We can use k-folds cross-validation. With a fixed value of λ, get the average value of errors. Take different values in fixed λ and do the same process and pick the best one. For example, we can grow the scale of *lambda* exponentially.

(4) LASSO Solution

Consider orthogonal design such that $X^TX = I$. Then $\hat{\beta}_j^{LASSO} = sgn(\hat{\beta}_j^{OLS})max(|\hat{\beta}^{OLS}| - \frac{\lambda}{2}, 0)$ When coefficients are small enough, they're truncated to be zeros. Otherwise, this method shirnks big coefficients by a constant $\frac{\lambda}{2}$.

(5) Consistency

LASSO has good properties in consistency. Let the true parameter vector be β_0 i.e. the true model $Y = \beta_0^T x + \varepsilon$ and the estimate be $\hat{\beta}_n$.

- Estimation Consistency
  $\hat{\beta}_n - \beta_0 \longrightarrow 0$ in p as $n \longrightarrow \infty$
- Model Selection Consistency
  Perform as a kind of feature selection (embedded method)
  $P(\{j : \hat{\beta}_j \neq 0\} = \{j : \beta_{0j} \neq 0\}) \longrightarrow 1$ as $n \longrightarrow \infty$
- Sign Consistency
  $P(sgn(\hat{\beta}_n) = sgn(\beta_0)) \longrightarrow 1$ as $n \longrightarrow n$
  This is stronger than model selection consistency.

More details are in [1] and [2].

12.3 Ridge Regression

(1) Idea

The difference with LASSO is that Ridge adds L2 norm instead of L1 norm. $\hat{\beta}_\lambda^{Ridge} = \min_\beta RSS + \lambda||\beta||_2^2$. Let us take a look at extreme cases. When $\lambda = 0$, $\hat{\beta}_\lambda^{Ridge} = \hat{\beta}^{OLS}$. When $\lambda \longrightarrow \infty$, $\hat{\beta}_\lambda^{Ridge} = 0$. The solution of ridge regression is $\hat{\beta}_\lambda^{Ridge} = (X^TX + \lambda I_p)^{-1}X^TY$.

Under high dimension ($p > n$), we can't say that X^TX has a full rank. So there may not exist an inverse of it. Hence $\hat{\beta}^{OLS}$ is not unique anymore. On the other hand, since $X^TX + \lambda I_p$ is invertible, the estimator from ridge method is decided uniquely.

References

[1] Asymptotics for lasso-type estimators, 2000, Fu, Knight

[2] On Model Selection Consistency of Lasso, 2006, Zhao, Yu