ch06_notes

August 10, 2019

Table of Contents

---

1 Linear Model Selection and Regularization

Alternatives to the least squares fitting procedures can yield better

- prediction accuracy
- model interpretability

1.1 Subset Selection

Methods for selecting a subset of the predictors to improve test performance

1.1.1 Best Subset Selection

Algorithm: Best Subset Selection (BSS) for linear regression

1. Let $\mathcal{M}_0$ denote the null model35
2. For $1 \leqslant k \leqslant p$:
    1. Fit all $\binom{p}{k}$ linear regression models with k predictors
    2. Let $\mathcal{M}_k = \underset{\text{models}}{\text{argmin}}\ RSS$
3. Choose the best model $\mathcal{M}_i, 1 \leqslant i \leqslant p$ based on estimated test error 36

For logistic regression, in step 2.A., let $\mathcal{M}_k = \underset{\text{models}}{\text{argmin}}\ D(y, \hat{y})$ where $D(y, \hat{y})$ is the ***deviance***37 of the model

Advantages

- Slightly faster than brute force. Model evaluation is $O(p)$ as opposed to $O(2^p)$ for brute force.
- Conceptually simple

Disadvantages

- Still very slow. Fitting is $O(2^p)$ as for brute force
- Overfitting and high variance of coefficient estimates when p is large

1.1.2 Stepwise Selection

Forward Stepwise Selection

Algorithm: *Forward Stepwise Selection* (FSS) for linear regression38

1. Let $\mathcal{M}_0$ denote the null model
2. For $0 \leqslant k \leqslant p - 1$:
    1. Fit all $p - k$ linear regression models that augment model $\mathcal{M}_k$ with one additional predictor
    2. Let $\mathcal{M}_{k+1} = \underset{\text{models}}{\text{argmin}}\ RSS$
3. Choose the best model $\mathcal{M}_i, 1 \leqslant i \leqslant p$ based on estimated test error

Advantages

- Faster than BSS. Fitting is $O(p^2)$ and evaluation is $O(p)$
- Can be applied in the high-dimensional setting $n < p$

Disadvantages

- Evaluation is more challenging since it compares models with different numbers of predictors.
- Searches less of the parameter space, hence may be suboptimal

Backward Stepwise Selection

Algorithm: Backward Stepwise Selection (BKSS) for linear regression 39 is

```
1. Let $\mathcal{M}_p$ denote the full model 40
2. For $k = p, p-1, \ldots, 1$:
    1. Fit all $k$ linear regression models of $k-1$ predictors that contain all but one of the predictors in $\mathcal{M}_k$.
    2. Let $\mathcal{M}_{k-1} = \underset{\text{models}}{\text{argmin}}\ RSS$
3. Choose the best model $\mathcal{M}_i, 1 \leqslant i \leqslant p$ based on estimated test error
```

Advantages

- As fast as FSS

Disadvantages

- Same disadvantages as FSS
- Cannot be used when $n < p$

Hybrid Approaches Other approaches exist which may add variables sequentially (as with FSS) but may also remove variables (as with BSS). These methods strike a balance between optimality (e.g. BSS) and speed (FSS/BSS)

1.1.3 Choosing the Optimal Model

Two common approaches to estimating the test error:

1. Estimate indirectly by adjusting the training error to account for overfitting bias
2. Estimate directly using a validation approach

C_p, AIC, BIC and Adjusted R^2

- Train MSE underestimates test MSE and decreases as p increases, so it cannot be used to select from models with different numbers of predictors. However we may adjust the training error to account for the model size, and use this to estimate the test MSE

- For least squares models, the C_p estimate41 of the test MSE for a model with d predictors is

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

where $\hat{\sigma} = \hat{\mathbb{V}}(\epsilon)$.

- For maximum likelihood models42, the ***Akaike Information Criterion*** (AIC) estimate of the test MSE is

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

- For least squares models, the ***Bayes Information Criterion*** (BIC) estimate43 of the test MSE is

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

- For least squares models, the ***adjusted*** R^2 statistic44 is

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

Validation and Cross-Validation

- Instead using adjusted training error to estimate test error indirectly, we can directly estimate using validation or cross-validation
- In the past this was computationally prohibitive but advances in computation have made this method very attractive.
- In this approach, we can select a model using the ***one-standard-error*** rule, i.e. selecting the model for which the estimated standard error is within one standard error of the p vs. error curve.

1.2 Shrinkage Methods

Methods for constraining or ***regularizing*** the coefficient estimates, i.e. ***shrinking*** them towards zero. This can significantly reduce their variance.

1.2.1 Ridge Regression

- Ridge regression introduces an L^2-penalty45 for the training error and estimates

  $$\hat{\beta}^R = RSS + \lambda\|\tilde{\beta}\|_2^2$$

  where λ is a ***tuning parameter***46 and $\tilde{\beta} = (\beta_1, \ldots, \beta_p)$47.

- The term $\lambda\|\beta\|_2^2$ is called a ***shrinkage penalty***

- Selecting a good value for λ is critical, see section 6.2.3

- Standardizing the predictors $X_i \mapsto \frac{X_i - \mu_i}{s_i}$ is advised.

Advantages

- Takes advantage of bias-variance tradeoff by decreasing flexibility 48 thus decreasing variance.
- Preferable to least squares in situations when the latter has high variance (close to linear relationship, $p \lesssim n$
- In contrast to least squares, works when $p > n$

Disadvantages

- Lower variance means higher bias.
- Will not eliminate any predictors which can be an issue for interpretation when p is large.

1.2.2 The Lasso

- Lasso regression introduces an L^1-penalty 49 for the training error and estimates

$$\hat{\beta}^R = RSS + \lambda \|\tilde{\beta}\|_1^2$$

Advantages

- Same advantages as ridge regression.
- Improves over ridge regression by yielding ***sparse models*** (i.e. performs variable selection) when λ is sufficiently large

Disadvantages

- Lower variance means higher bias.

Another Formulation for Ridge Regression and the Lasso

- Ridge Regression is equivalent to the quadratic optimization problem:

$$\begin{aligned} \min RSS + \|\tilde{\beta}\|_2 \\ \text{s.t. } \|\tilde{\beta}\|_2^2 \leqslant s \end{aligned}$$

- Lasso Regression is equivalent to the quadratic optimization problem:

$$\begin{aligned} \min RSS + \|\tilde{\beta}\|_1 \\ \text{s.t. } \|\tilde{\beta}\|_1 \leqslant s \end{aligned}$$

Bayesian Interpretation for Ridge and Lasso Regression Given Gaussian errors, and simple assumptions on the prior $p(\beta)$, ridge and lasso regression emerge as solutions

- If the $\beta_i \sim \text{Normal}(0, h(\lambda))$ iid for some function $h = h(\lambda)$ then the ***posterior mode*** for β (i.e. $\underset{\beta}{\text{argmax}} p(\beta|X, Y)$) is the ridge regression solution
- If the $\beta_i \sim \text{Laplace}(0, h(\lambda))$ iid then the posterior mode is the lasso regression solution.

1.2.3 Selecting the Tuning Parameter

Compute the cross-validation error $CV_{(n),i}$ for for a "grid" (evenly-spaced discrete set) of values λ_i, and choose

$$\lambda = \underset{i}{\operatorname{argmin}}\, CV_{(n),i}$$

1.3 Dimension Reduction Methods

- ***Dimension reduction*** methods transform the predictors $X_1, \ldots, X_p$ into a smaller set of predictors $Z_1, \ldots, Z_M$, $M < p$.
- When $p >> n$, $M << p$ can greatly reduce the variance of the coefficient estimates.
- In this section we consider linear transformations

  $$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

  and a least squares regression model

  $$Y = \mathbf{Z}\theta + \epsilon$$

  where $\mathbf{Z} = (1, Z_1, \ldots, Z_M)$

1.3.1 Principal Components Regression

Principal Components Analysis is a popular unsupervised approach 50 that can be used for dimensional reduction

An Overview of Principal Components Analysis

- The ***principal components*** of a data matrix $n \times p$ matrix $\mathbf{X}$ can be seen (among many different perspectives) as the right singular eigenvectors $v_1, \ldots, v_p$ of the $p \times p$ sample covariance matrix C, i.e. the eigenvectors of $C^\top C$) ordered by decreasing absolute value of the corresponding eigenvalues.
- Let $\sigma_1^2, \ldots, \sigma_k^2$ be the singular values of C (the squares of the eigenvalues of $C^\top C$) and let $v_1, \ldots, v_p$ be the corresponding eigenvectors of C. Then σ_i^2 is the variance of the data along the direction v_i, and σ_1^2 is the direction of maximal variance.

The Principal Components Regression Approach

- ***Principal Components Regression*** takes $Z_1, \ldots, Z_M$ to be the first M principal components of $\mathbf{X}$ and then fits a least squares model on these components.
- The assumption is that, since the principal components correspond to the directions of greatest variation of the data, they show the most association with Y. Furthermore, they are ordered by decreasing magnitude of association.
- Typically M is chosen by cross-validation.

Advantages

- If the assumption holds then the least squares model on $Z_1, \dots, Z_M$ will perform better than $X_1, \dots, X_p$, since it will contain most of the information related to the response 51, and by choosing $M << p$ we can mitigate overfitting.

- Decreased variance of coefficient estimates relative to OLS regression

Disadvantages

- Is not a feature selection method, since each Z_i is a linear function of the predictors

Recommendations

- Data should usually be standarized prior to finding the principal components.

1.3.2 Partial Least Squares

A supervised dimension reduction method which proceeds roughly as follows

- Standardize the variables
- Compute Z_1 by setting $\phi_{j1} = \hat{\beta}_j$ the ordinary least squares estimate 52
- For $ 1 < m < M$, Z_m is determined by
  - Adjust the data $X_j = \epsilon_j$ where ϵ_j is the residual from regression of Z_{m-1} onto X_j
  - Compute Z_m in the same fashion as Z_1 on the adjusted data

As with PCR, M is chosen by cross-validation

Advantages

- Decreased variance of coefficient estimates relative to OLS regression
- Supervised dimension reduction may reduce bias

Disadvantages

- May increase variance relative to PCR (which is unsupervised).
- May be no better than PCR in practice

1.4 Considerations in High Dimensions

1.4.1 High-Dimensional Data

Low dimensional means $p << n$, high dimensional is $p \gtrsim n$

1.4.2 What Goes Wrong in High Dimensions?

- If $p \gtrsim n$, then linear models will create a perfect fit, hence overfit (usually badly)

- C_p, AIC, BIC, and R^2 approaches don't work in well in this setting

1.4.3 Regression in High Dimensions

1. Regularization or shrinkage plays a key role in high-dimensional problems.
2. Appropriate tuning parameter selection is crucial for good predictive performance.
3. The test error tends to increase as the dimensionality of the problem increases if the additional features aren't truly associated with the response (the curse of dimensionality)

1.4.4 Interpreting Results in High Dimensions

- Multicollinearity problem is maximal in high dimensional setting
- This makes interpretation difficult, since models obtained from highly multicollinear data fail to identify which features are "preferred"
- Care must be taken to measure performance 53

1.5 Footnotes

35. This is the model that predicts $\hat{y} = \overline{y}$, i.e. $\hat{\beta}_i = 0$ for $i > 1$ and $\hat{\beta}_0 = \overline{y}$.

36. Estimates of test error can come from CV, $C_p(AIC)$, BIC or adjusted R^2

37. Here $D(y, \hat{y}) = -2\log(p(y \mid \hat{\beta})$ where $\hat{\beta}$ is the MLE for β.The author's definition of deviance can be found in the comment on the Wikipedia entry if $\hat{\theta}_0$

38. As with BSS, we can use FSS for logistic regression by replacing RSS with the deviance in step 2B.

39. As with BSS, we can use BackSS for logistic regression by replacing RSS with the deviance in step 2B.

40. Here full means contains all p predictors.

41. Thus C_p is RSS plus a penalty which depends on the number of predictors and the estimate of the error variance. One can show that if $\hat{\sigma}^2$ is unbiased then then C_p is unbiased.

42. For Gaussian errors, the least squares estimate is the maximumlikelihood estimate so in that case C_p and AIC are proportional.

43. The BIC places a heavier penalty than C_p when $n > 7$ due to the $\log(n)d\hat{\sigma}^2$ term. The book says this means BIC places a heavier penalty than C_p on models with many variables although this isn't clear. It would seem it places a penalty on large numbers of observation (unless somehow larger numbers of observations are correlated with larger numbers of predictors).

44. C_p, AIC and BIC are all estimates of the test MSE so smaller values are better. By contrast, larger values of adjusted R^2, but this is equivalent to minimizing $RSS/(n - d - 1)$ which likely can be thought of as a test MSE estimate.