6 Linear Model Selection and Regularization

Andrew Liang

10/7/2020

Notes

Subset Selection

Best Subset

1. Fit M_0, the null model, with no predictors. (only predicts sample mean for each observation).
2. For $k = 1, 2, \ldots, p$:
    - Fit all $\binom{p}{k}$ models that contain exactly k predictors
    - Choose the best among the $\binom{p}{k}$ models and call it M_k. Best is defined as having smallest RSS, or equivalently largest R^2
3. Select single best model among $M_0, \ldots, M_p$ using CV prediction error, $C_p(AIC)$, BIC, or adjusted R^2

- Suffers from computational limitations, as the number of possible models grows rapidly as p increases (2^p models)

Forward Stepwise Selection

1. Fit M_0, the null model, with no predictors.
2. For $k = 0, \ldots, p-1$:
    - Consider all $p-k$ models that augment the predictors in M_k with one additional predictor
    - Choose best among $p-k$ models (M_{k+1})
3. Select single best model among $M_0, \ldots, M_p$ using CV prediction error, $C_p(AIC)$, BIC, or adjusted R^2

- Much less computationally expensive compared to best subset
- However, not guaranteed to find best subset model
- Can be applied in high-dimensional setting ($n < p$)

Backward Stepwise Selection

1. Fit M_p, the full model, with all predictors.
2. For $k = p, p-1, \ldots, 1$:
    - Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors
    - Choose best among k models (M_{k-1})
3. Select single best model among $M_0, \ldots, M_p$ using CV prediction error, $C_p(AIC)$, BIC, or adjusted R^2

- Also not guaranteed to find best model
- REQUIRES that n is larger than p

Best subset, forward, and backward selection generally give similar but not identical models

Choosing the Optimal Model

Techniques for adjusting the training error for the model size are available

1. C_p
    - for a fitted least squares model containing d predictors and the variance of the error $\hat{\sigma}^2$, C_p estiamte of test MSE is:
    $$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$
    - penalty increases as number of predictors in model increases
    - choose model with lowest C_p value
2. AIC
    - defined for models fit by maximum likelihood (least squares)
    $$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$
    - proportional to C_p
3. BIC (similar to C_p and AIC, but from a Bayesian POV)
    $$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2log(n)d\hat{\sigma}^2)$$
    - replaces $2d\hat{\sigma}^2$ with $log(n)d\hat{\sigma}^2$
    - since $log(n) > 2$ for any $n > 7$, BIC generally places heavier penalty on models with many predictors
4. Adjusted R^2
    $$AdjustedR^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$
    - unlike previous penalties, we want to choose model with highest adjusted R^2
    - despite popularity, is not as statistically motivated as the previous penalties

Shrinkage Methods

- fit model using all predictors and regularizes coefficients/shrinks coefficients towards zero
    - reduces variance

Ridge Regression

wants to minimize:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

- $\lambda\sum_{j=1}^{n}\beta_j^2$ is the shrinkage penalty
- $\lambda \geq 0$ is the tuning parameter
  - as $\lambda \to \infty$, the model coefficients approaches zero (except for model intercept β_0)
- selecting λ value is important (can use CV)
- best to apply ridge after predictors have been standardized (due to potential scaling issues):

$$\tilde{x_{ij}} = \frac{x_{ij}}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2)}}$$

- important to note that all the predictors will still be included in the model; only the magnitude of the coefficients is affected

The Lasso

- similar to ridge, but has the ability to exclude predictors in final model (better for interpretability)

wants to minimize:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

- λ penalty has the effect of forcing some of the coefficient estimates to be zero when λ is sufficiently large

Ridge vs Lasso

- generally, ridge performs better when response is a function of many predictors, with all coefficients roughly the same size
- generally, lasso performs better when only a relatively small number of predictors have substantial coefficients, and remaining variables are very small coefficients
- both perform shrinkage, whereas ridge shrinks the coefficients by the same proportion, whereas lasso shrinks all coefficients toward 0 by the same amount, and sufficiently small coefficients are shrunken all the way to 0

Dimension Reduction Methods

- idea is to transform the predictors then fit a least squares model

let $Z_1, Z_2, \cdots, Z_M$ represent $M < p$ linear combinations of original p predictors:

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

for some constants $\phi_{1m}, \phi_{2m}, \cdots, \phi_{pm}$, then we fit the linear regression model:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i$$

* dimension of the problem has been reduced from $p+1$ to $M+1$ * can often outperform least squares IF the choice of $Z_1, Z_2, \cdots, Z_M$ is chosen wisely

Principal Components Analysis (PCA)

- dimension reduction technique in which the *first principle component* direction of the data is that along which the observations *vary the most* (have highest variance)
  - is a vector that defines a line that minimizes perpendicular distances between each point and the line (distance represents the projection of the point onto that line)
- PCA scores for the 1st component is defined as:

$$Z_{j1} = \sum_{j=1}^{p} \beta_j (X_j - \overline{X}_j)$$

* can calculate up to p distinct pcincipal components * 2nd PC is a linear combination of variables that is uncorrelated with Z_1, or equivalently must be perpendicular/orthogonal to Z_1 * first component will always contain the most info

Principal Components Regression Approach (PCR)

- involved using $Z_1, Z_2, \cdots, Z_M$ as predictors in linear regression
- assume that the directions in which $X_1, \ldots, X_p$ *show the most variation are the directions that are associated with* Y
- will be better than the original linear model with $X_1, \ldots, X_p$ as predictors if PCR assumptions are met
- performs better when the first few principal components are sufficient to capture most of variation in the predictors and their relationships with the response
- since PCR is a lienar combination of all p of the *original* features, it is not a feature selection method
- number of components M usually chosen by CV
- usually recommended to standardize predictors using method from ridge if these predictors aren't on the same scale
- example of an *unsupervised* method

Partial Least Squares (PLS)

- a supervised method similar to PCA where it is dimension reduction
- same process as PCR, but also uses response Y to find directions that help explain both response and predictors
  - places highest weight on variables strongly correlated with Y
- often performs no better than PCR or ridge

Considerations in High Dimensional Data

- when $p \geq n$, linear regression/logistic regression should not be performed
- C_p, AIC, BIC unfortunately are not appropriate in high dimensional settings, as estimating $\hat{\sigma}^2$ is problematic
- 3 important points:

    1. regularization/shrinkage is very important in high-dimensional settings
    2. appropriate tuning parameter selection key for good predictive performance
    3. test error tends to increase as dimensionality increases, unless the additional predictors are truly associated with response

- adding new features is a truly a double-edged sword, depending whether or not they are truly associated with Y
- should *never* use sum of squared errors, p-values, R^2 statistics as evidence of model fit in high dimensional setting

Applied