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:

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

- $\lambda \sum_{j=1}^{n} \beta_j^2$ 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):

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