DSO530 Statistical Learning Methods

Lecture 6 part I: Linear Model Selection

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California

Outline

- This lecture is based on the first part of Chap. 6 in `ISLR`
- Previously, we studied the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon\,.$$

- One typically fits this model using least squares (go back to Lecture 2. if you do not recall the concept).
- We might want to use another fitting procedure instead of least squares to yield better prediction accuracy and model interpretability
- In this lecture, we discuss one alternative approach to using least squares to fit linear models: subset selection.
- This set of slides only covers the subset selection methods ideas. A `Python` tutorial will be released later.

Subset Selection

- Subset selection methods include best subset selection, stepwise selection.
- To perform *best subset selection*, we fit a separate least squares regression for each possible combination of the p predictors
- Potential problem: selecting the best model from among too many possibilities considered by best subset selection is not trivial.
- Best subset selection is usually implemented by:

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

A few questions on best subset selection

- In step 2b), why can we just look at the RSS on training error?
- In step 3), why don't we just look at the RSS on training error?
- How many models do we search through in best subset selection?

Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Figure 1

Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Figure 2

- In either forward or backward selection, we search through $1 + p(p+1)/2$ models. This is a huge saving compared with best subset selection
- However, forward stepwise selection and backward stepwise selection might miss the optimal subset of features. This is a price we have to pay for computational advantages.
- (Optional) There are hybrid approaches that combine forward and backward selection (p. 210 of `ISLR`)

Choose the model in Step 3)

In best subset, forward and backward selection algorithms, the step 3)'s are the same. There are essentially two ideas in this step

- *directly* estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in Chapter 5. A variant of the cv approach is the one-standard-error rule.
- *indirectly* estimate test error by making an adjustment to the training error (e.g., adjusted R^2, AIC, BIC and C_p).
- $$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

  Q: how is adjusted R^2 different from R^2?
- Q: The d is clear in least squares regression. But what if we have some restrictions on the variable coefficients?

C_p, AIC and BIC (for linear regression)

- Mallow's C_p

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

- Akaike information criterion (AIC)

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

- Bayesian information criterion(BIC)

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

- Note that $\hat{\sigma}^2$ is the estimated variance of random error term ϵ using the full model (i.e., p predictors).
- The definitions of AIC and BIC ignored some constant

A few questions

- For linear regression, will C_p and AIC give the same ranking of models?
- When n is bigger than 7, $\log n > 2$. This means BIC penalizes larger models heavier compared to AIC. So which criterion encourages smaller models?
- In the definitions of adjusted R^2, AIC, BIC and C_p, do you see the trade-offs between fitting on training data and model complexity?
- When $p > n$, estimating $\hat{\sigma}^2$ is a big problem. Then which method do you prefer for model selection?
- Among adjusted R^2, AIC, C_p, BIC and cross-validation, which one is the most easily generalizable beyond least squares linear regression?