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
xint@marshall.usc.edu

Outline

- Previously, we studied the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

- 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 on all p features to yield better prediction accuracy and model interpretability
- In this lecture, we discuss one alternative approach: subset selection.
- Subset selection methods include best subset selection, stepwise selection.
- This set of slides only covers the subset selection methods ideas. A `Python` tutorial will be released later.

Best Subset Selection

a) $p!$ b) 2^p c) Neither A or B

$\{x_1, \ldots, x_p\}$ How many subsets are there?

$2 \times 2 \ldots \times 2$

$M_0 = \{\} = 1 = {}^pC_0$

$M_1 = \{x_1\} \ldots \{x_p\} = {}^pC_1$

$M_2 = \{x_1, x_2\} \ldots \{x_{p-1}, x_p\} = {}^pC_2$

$\vdots$

$M_p = \{x_1 \ldots x_p\} = 1 = {}^pC_p$

$${}^pC_0 + {}^pC_1 + \ldots + {}^pC_p = 2^p$$

e.g. All subsets: $\{x_1, x_2\}$

$\{\}, \{x_1\}, \{x_2\}, \{x_1, x_2\}$

if $p = 20$

$2^p = 2^{20}$ subsets

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

$${}^pC_r = {}^pC_{p-r}$$

This is just enumeration, not selection.

1 option: choose best set at each level of highest R^2 or smallest RSS

Since models are of same size, in sample R^2 can be used.

Best Subset 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.

Q If we stop the best subset at k^* ($< p$) then how many models?
a) 2^k b) $k^*!$ c) Neither A or B

A few questions on best subset selection

$p=2$ $k^{\#}=1$ $\{x_1, x_2\}$ subsets : $\{\}$ $\{x_1\}$ $\{x_2\}$

Ans] $P_{C_0} + P_{C_1} + \ldots + P_{C_{k^{\#}}}$

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

models of same size

different size

$p = 20$

$2^{20} > 1,000,000$

$p^2 = 400$

Forward Stepwise Selection

⇒ As subset selection can be slow sometimes

M_0 : {}

M_1 : $\{x_1\}$ $\{x_2\}$ ... $\{x_p\}$ p

M_2 : $\{x_1, x_2\}$ ... $\{x_1, x_p\}$ p-1 models

M_3 : ... p-2

: ...

M_s : {...} 1

this is where forward selection & subset method differ. e.g. x_1 is always there is stepwise imp

'date subset depends on previous level.

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.

∴ total # models : $1 + p + (p-1) + \ldots + 1 = 1 + \frac{p(p+1)}{2}$

$= O(p^2)$

p = 20

$2^p > 1{,}000{,}000$

$p^2 = 400$.

potential problem: some best subsets may be left out

Q. If we stop fss at $k^*(p)$ then how many models?

Backward Stepwise Selection

M_p : $x_1 \ldots x_p$ } total 1

M_{p-1} : $\{x_2, x_3 \ldots x_p\}, \{x_1, x_3 \ldots x_p\}, \ldots \{x_1 \ldots x_{p-1}\}$ total p

M_{p-2} : $\{x_3 \ldots x_p\}, \{x_2, x_4 \ldots x_p\}, \ldots \{x_2, x_3 \ldots x_{p-1}\}$

$\vdots$

$M_0 = \{\}$ total 1

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.

- 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.

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 Lecture 5.
- *indirectly* estimate test error by making an adjustment to the training error (e.g., adjusted R^2, AIC, BIC and C_p).
- $$\textit{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

  Q: how is adjusted R^2 different from R^2?

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)

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

trade off bw model fitting & complexity

- 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

B

- For linear regression, will C_p and AIC give the same ranking of models? Yes.
- 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? IC

A few questions

- for $p > n$, backward X.
- we can use best subset & forward w early stopping at k^* (CV)
- CV is best approach.

- 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?