Subset Methods

Wednesday, September 22, 2021 6:08 PM

- Subset Selection
  - Best Subset
    - Algorithm:
      1. M_0 is null model (no features)
      2. For $k = 1 \ldots p$
         a. Fit all $\binom{p}{k}$ models with k predictors
         b. Select best of the k models, called M_k
            ↑ lowest RSS, highest R^2...
      3. Select best model overall from $M_0 \ldots M_p$
         using AIC / BIC / Adj. R^2

    - Computational Considerations / Cost
      - Brute Force
      - Intractable for large p

  - Stepwise Selection
    - Algorithm (Forward)
      1. M_0 as null model
      2. For $k = 0 \ldots p-1$
         a. Consider all $p-k$ models that augment the predictors in M_k with one additional predictor
         b. Choose best of $p-k$ models → M_{k+1}
            ↑ lowest RSS, highest R^2
      3. Select best model from $M_0 \ldots M_p$
         ↑

~ AIC / BIC / Adj R^2

* Greedy Strategy
* If $n < p$ in orig sample ⇒ Forward Selection

- Algorithm (Backward)
  1. M_0 as null model
  2. For $k = p \dots 1$
     a. Consider all k models that contain all but one of the predictors in M_k for $k-1$ predictors
     b. Choose best among k models ⇒ M_{k-1}
        ↑ lowest RSS, highest R^2
  3. Select best model $M_0 \dots M_p$
     ↑ AIC / BIC / Adj R^2

* Greedy, requires $n > p$

- Comparison of $M_0 \dots M_k$ (Optimal selection)
  - Given multiple models ⇒ Directly estimate $E[\text{error}]$
  - Indirectly adjust scores for multiple models
    ⇒ Penalize # of features!

- AIC (C_p)

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

↑ penalty of # feat × quality

(error)

- BIC

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

- Adjusted R^2

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

$$\max Adj\ R^2 \Rightarrow \min \frac{RSS}{n-d-1}$$