QTM 347 Machine Learning

Lecture 9: Subset selection

Ruoxuan Xiong

Suggested reading: ISL Chapter 6

Lecture plan

- Best subset selection

Motivation

- In many modern datasets, the **dimension** of feature vector is **much larger than** the **number of examples** in the training set
  - *High-dimensional datasets*
  - A canonical example is *biological gene expressions*: $p \gg n$

Features/Input $\quad x = \begin{array}{c} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_p \end{bmatrix} \end{array} \begin{array}{l} \text{--- gene 1} \\ \text{--- gene 2} \\ \text{--- gene 3} \\ \\ \vdots \\ \text{--- gene } p \end{array} \quad \longrightarrow \quad y$ Label/Output

- **Overfitting** is prevalent when learning from *high-dimensional datasets*

Subset selection

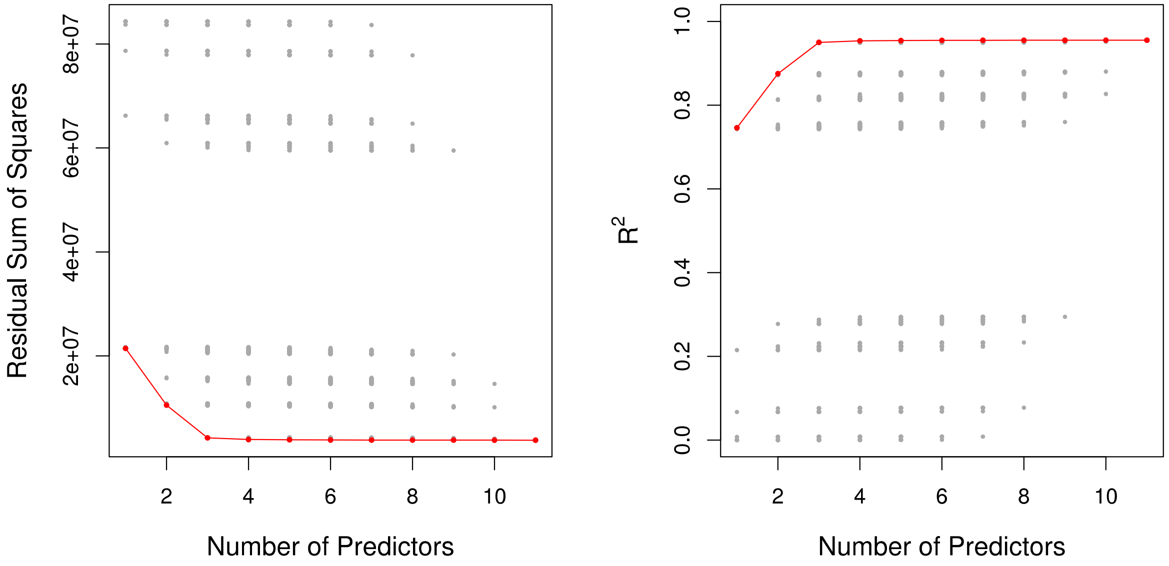
- **Step 1**: For each k, select a subset of k predictors from the total $\boldsymbol{p}$ predictors
  - There are $\binom{p}{k} = \frac{p!}{k!(p-k)!}$ possible ways of choosing $\boldsymbol{k}$ predictors
  - Choose the subset with the smallest residual sum of squares (or whichever loss we decide to use)
- **Step 2**: Select the optimal k

Example

- Credit card data set
  - Predict whether customers default on their credit card debt
  - Predictors (11 predictors in total)
    - Income: Income in $1,000's
    - Limit: Credit limit
    - Rating: Credit rating
    - Cards: Number of credit cards
    - Age: Age in years
    - Education: Number of years of education
    - Gender: A factor with levels Male and Female
    - Student: A factor with levels No and Yes indicating the individual was a student
    - Married: A factor with levels No and Yes indicating whether the individual was married
    - Ethnicity: A factor with levels African American, Asian, and Caucasian indicating the individual's ethnicity
    - Balance: Average credit card balance in $

Example

- Best model for a given number of predictors



- Both residual sum of squares and R^2 improve as we increase k
- We do not want select the model with maximum number of predictors

Objective in subset selection

- Select the optimal k and the set of k predictors to minimize the test error

- **Cross-validation** is one approach to **directly** estimate the test error

- **Alternative criteria** to **indirectly** estimate the test error by making an **adjustment** to the **training error**
  - Less expensive to compute
  - Computational cost matters in subset selection since we are looking at $\binom{p}{k}$ many combinations of predictors

Three alternative criteria

- Akaike Information Criterion (AIC)

- Bayesian Information Criterion (BIC)

- Adjusted R^2 statistic

Model selection criteria I: AIC

$$C_p = \frac{1}{n}(\text{RSS} + 2k\hat{\sigma}^2)$$

- $\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is predicted label. $\text{MSE}_{\text{training}} = \frac{1}{n}\text{RSS}$
- $\hat{\sigma}^2$ is an estimate of the variance of the "error" using the full model containing all predictors
- k is the number of predictors in the model

- **Interpretation**
  - Adds a penalty of $2k\hat{\sigma}^2$
  - Penalty increases as k---number of predictors in the model---increases

Model selection criteria II: BIC

$$\frac{1}{n}\left(\mathrm{RSS} + \log(n)\, k\hat{\sigma}^2\right)$$

- $\hat{\sigma}^2$ is an estimate of the variance of the "error"
- k is the number of predictors in the model

- **Interpretation**
  - Adds a penalty of $\log(n)\, k\hat{\sigma}^2$
  - When $\log(n) > 2$ (or $n > 7$), BIC places a heavier penalty on having more predictors than AIC

Model selection criteria III: Adjusted R^2 statistic

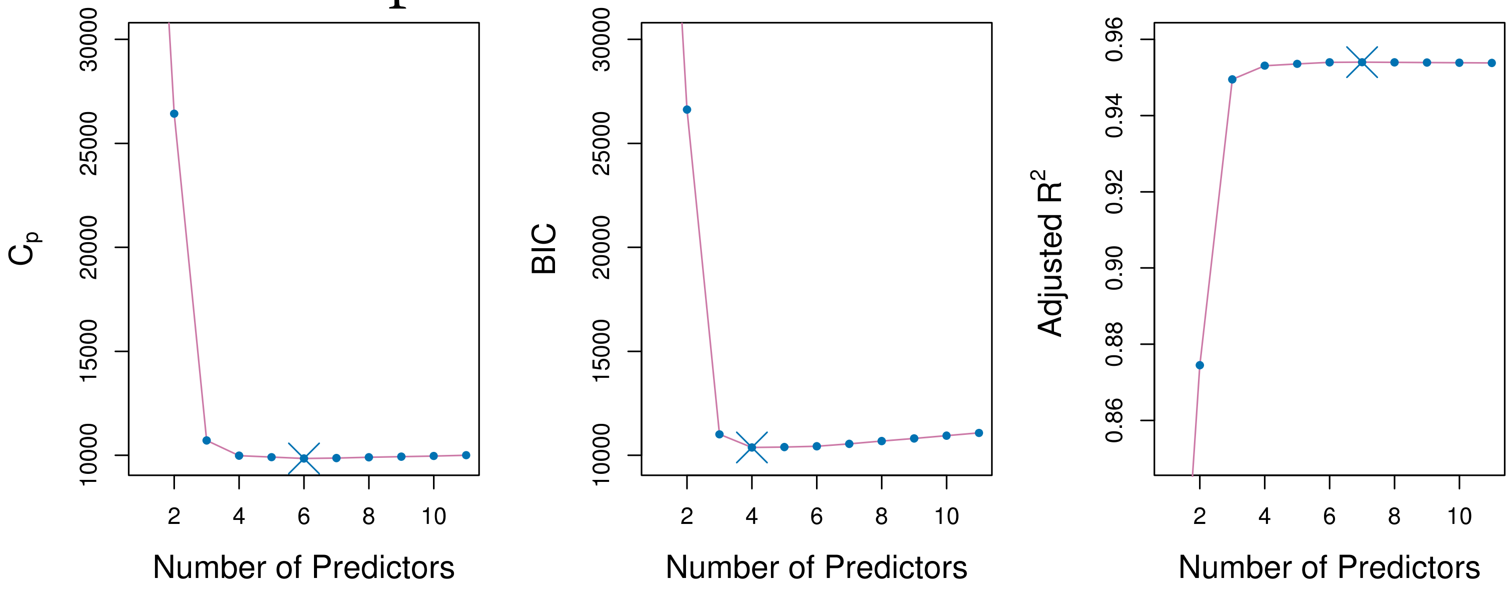
$$R^2_{\mathrm{adj}} = 1 - \frac{\frac{\mathrm{RSS}}{n-k-1}}{\frac{\mathrm{TSS}}{n-1}}$$

- $\mathrm{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- $\mathrm{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- $R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$
- R^2_{adj} may not always increase with k

- **Interpretation**
  - Once all of the "correct" variables have been included, adding additional "noise" variables will lead to only a *small* decrease in RSS

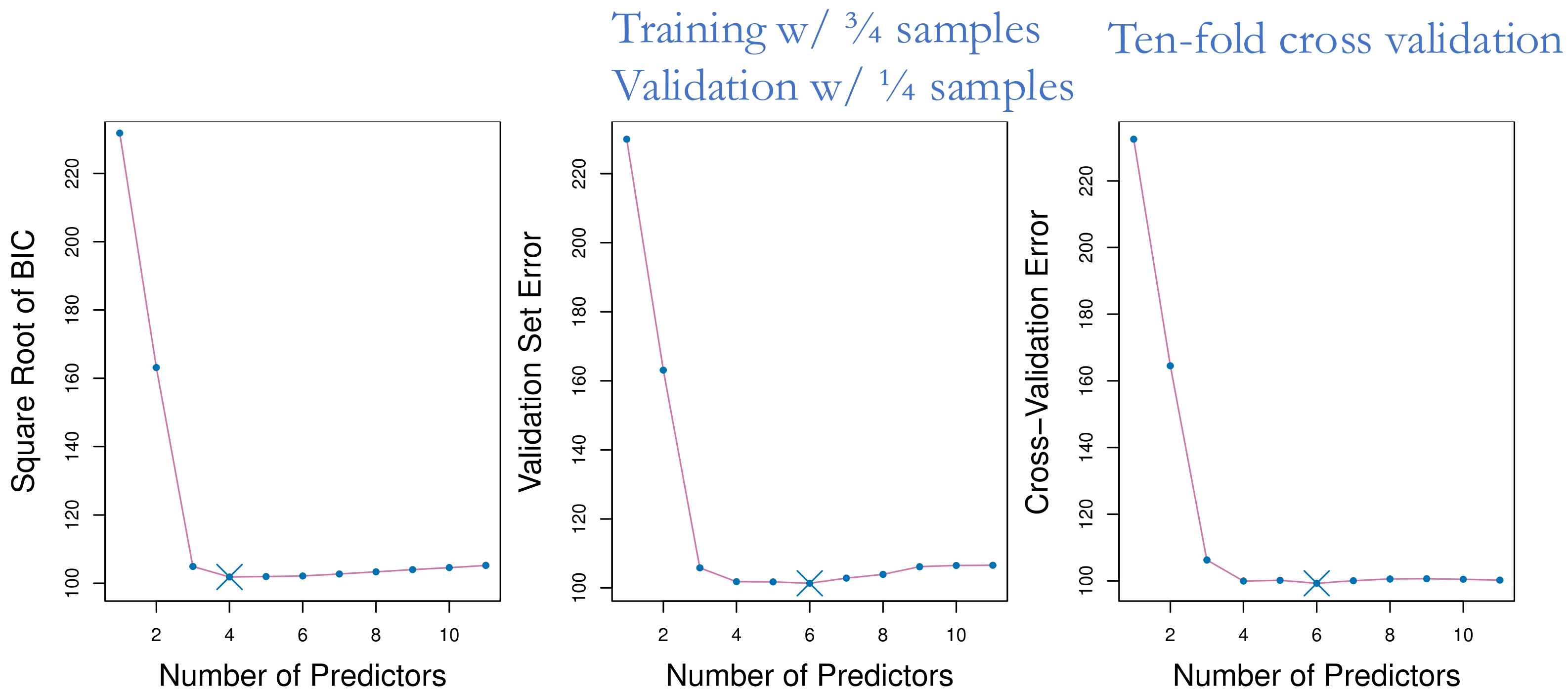
Comparison between model selection criteria

- **Example**: Best subset selection for the credit card data set
- Criteria vs. number of predictors:



- BIC selects a model with the smallest number of predictors
- Recall that when $\log(n) > 2$ or $n > 7$, BIC places a heavier penalty on models with more predictors than AIC

Comparison with cross-validation



- **Recall**: In k-fold cross validation, estimate standard error for test error
  - **One standard error rule:** simplest model within one standard error from lowest
  - Select a 3- or 4-variable model according to this rule

Cross-validation vs. Evaluation criteria

- **Cross-validation** is computationally expansive
  - Provides a **direct** estimate of the test error
  - Make **fewer** assumptions on the true model

- **These evaluation criteria (AIC, BIC, adjusted R^2)** are computationally cheap
  - Only works under certain assumptions on the true model

Discussions

- Best subset selection has two problems:
  - Computationally expensive: fit over 2^p models!
  - Too many possibilities increases chances of overfitting
    - Selected model has *high variance*

- **Possible solution**: restrict search space

- **Next**: Stepwise selection methods