# Lecture 5 Model Selection I

## ECE 625: Data Analysis and Knowledge Discovery

Di Niu

Department of Electrical and Computer Engineering
University of Alberta

January 26, 2021

# Outline

# Why Model Selection

- In many situations, many predictors are available. Some times, the number of predictors is even larger than the number of observations ($p > n$). We follow Occam's razor (aka Ockham's razor), the law of parsimony, economy, or succinctness, to include only the important predictors.
- The model will become simpler and easier to interpret (unimportant predictors are eliminated).
- Cost of prediction is reduced-there are fewer variables to measure.

  Generalizability improves
- Accuracy of predicting new values of $y$ may improve.
- Recall MSE(prediction) = Bias(prediction)$^2$ + Var (prediction).
- Variable selection is a tradeoff between the bias and variance.

  by tuning p, which indicates the flexibility

# How to select model in Linear Regression

- **Subset Selection.** We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables. Best subset and stepwise model selection. based on some statistics
- **Shrinkage.** We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.
- **Dimension Reduction.** We project the p predictors into a $M$-dimensional subspace, where $M < p$. This is achieved by computing $M$ different linear combinations, or projections, of the variables. Then these $M$ projections are used as predictors to fit a linear regression model by least squares.

# Training Error vs. Testing Error Estimating Generalization Gap

- Let $\mathcal{T} = \{(x_1, y_1), ...(x_N, y_N)\}$ be a training set. Training error:

$$\overline{\text{err}} = \frac{1}{N} \sum_{i}^{N} L(y_i, \hat{f}(x_i)).$$

- Given $x_i$, the response variable $Y_i$ is a random variable. The *in-sample* error is

estimate of the test error

yi^0 = f(xi)+ei

$$\text{Err}_{\text{in}} = \frac{1}{N} \sum_{i=1}^{N} E_{Y_i^0}[L(Y_i^0, \hat{f}(x_i))|\mathcal{T}],$$

reused

where $Y_i^0$ is a new response at $x_i$. $\text{Err}_{\text{in}}$ is a good estimate of the testing error on other samples. Then for squared loss (Chapter 7) we have

$$E_{\mathbf{y}}[\text{Err}_{\text{in}}] = E_{\mathbf{y}}[\overline{\text{err}}] + \frac{2}{N} \sum_{i}^{N} \text{Cov}(\hat{y}_i, y_i) = E_{\mathbf{y}}[\overline{\text{err}}] + 2 \cdot \frac{d}{N} \sigma_\epsilon^2,$$

generalization gap

where the 2nd equality holds if $\hat{y}_i$ is a linear fit with $d$ inputs.

# Best subset selection

- Fit all possible models ($2^p - 1$) and select a single best model from according certain criteria.
- Possible criteria include $C_p$, AIC, BIC, adjusted $\mathrm{R}^2$, or cross-validated prediction error.
- The adjusted $R^2$ statistic:

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

  where $d$ is the number of predictors in the model.
- $R^2$ is NOT suitable for selecting the best model as it always select the largest model to have smallest training error while we need to have small testing error.
- Adjusted $R^2$ criterion: we pick the best model by maximizing the adjusted $R^2$ over all $2^p - 1$ models.

# $C_p$ Statistic

- The $C_p$ statistic is another statistic which penalizes larger model:

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

- Mallow's $C_p$ statistic is $C'_p = \text{RSS}/\hat{\sigma}^2 + 2d - n$, consistent with above.
- It can be shown that $C'_p \approx d + 1$, if all the important predictors are in the model.
- $C_p$ criterion: pick the model such that $C_p(d)$ is close to $d + 1$ and also $d$ is small (we prefer simpler models).

# AIC Criterion

- The AIC statistic for a model is defined by maximum likelihood.

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

  where $d$ is the number of predictors in the model.
- In linear model, under Gaussian error, $C_p$ is proportional to AIC.
- AIC criterion: pick the best model by minimizing AIC criterion over all models.

# BIC Criterion

- The BIC statistics for a model is defined as

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

  where $d$ is the number of predictors in the model.
- Similar to AIC, the BIC has the second part to penalize larger models.
- But compared to AIC, BIC tends to select even smaller models due to $\log(n)$.
- BIC criterion: pick the best model by minimizing BIC criterion over all models.
- The BIC criterion can guarantee that we can pick all the important predictors as $n \longrightarrow \infty$, while the AIC criterion cannot.

# Cross-Validation

y1, y2, …~~yi~~, …..yn

Train->\hat beta_{-i}
Use \hat beta_{-i} to predict y_i —>\hat y_{-i}

- The idea of cross-validation (CV) criterion is to find a model which minimizes the prediction/testing error.

Leave-one-out CV (LOOCV)

- For $i = 1, \ldots, n$, delete the $i$-th observation from the data and the linear regression model. Let $\hat{\boldsymbol{\beta}}_{-i}$ denote the LSE for $\boldsymbol{\beta}$. Predict $y_i$ using $\hat{y}_{-i} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{-i}$.
- CV criterion: pick the best model by minimizing the $\text{CV} = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2$ statistic over all the models.
- We did not use $y_i$ to get $\hat{\boldsymbol{\beta}}_{-i}$ and we predict $y_i$ as if it were new "observation".

CV is another way to estimate the generalizability of your model
It's nothing about the training error,
but about how well a model could do on data it has never seen.

# Backward Elimination A Greedy Algorithm

- Backward elimination starts with all $p$ predictors in the model. Delete the least significant predictor.
- Fit the model containing all the $p$ predictors $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$ and for each predictor calculate the $p$-value of $t$-test (the single $F$-test). Other criteria, say, AIC, BIC, and $C_p$ apply as well.
- Check whether the $p$-values for all the $p$ predictors are smaller than $\alpha$, called alpha to drop.
- If yes, stop the algorithm and all the $p$ predictors are treated as important.
- If not, remove the least significant variable, i.e., the variable with the largest $p$-value and repeat checking.

# Forward Selection A Greedy Algorithm

- Forward Selection starts with no predictor in the model. Pick the most significant predictor.
- Fit $p$ simple linear regression models

$$y = \beta_0 + \beta_1 x_j, \;\; j = 1, \ldots, p.$$

For each predictor, we calculate the $p$-value of the $t$-test for the hypothesis $H_0 : \;\beta_1 = 0$. Other criteria, say, AIC, BIC, and $C_p$ apply as well.
- Choose the most significant predictor from the remaining predictors, denoted by $x_{(1)}$ such that the $p$-value for the hypothesis $H_0 : \;\beta_1 = 0$ is smallest.
- If the $p$-value for the most significant predictor is larger than $\alpha$ (alpha to enter). We stop and no more predictor is needed.
- If not, the most significant predictor is added in the model and we repeat choosing.

# Stepwise selection

Still a greedy algorithm

- A disadvantage of backward elimination is that once a predictor is removed, the algorithm does not allow it to be reconsidered.
- Similarly, with forward selection once a predictor is in the model, its usefulness is not reassessed at later steps.
- Stepwise selection, which is a hybrid of the backward elimination and the forward selection, allows the predictors enter and leave the model several times.
- Forward stage: Do Forward Selection until stop. until nothing can be included in the model
- Backward stage: Do Backward Elimination until stop. until nothing should be eliminated
- Alternate between the above two stages until no predictor can be added and no predictor can be removed according to the specified $\alpha$ to enter and $\alpha$ to drop.

# Summary and Remark

- Introduction
- Best subset selection
- Stepwise method
- Read textbook Chapter 3
- Do R lab