UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Lecture 5 Model Selection I

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Outline

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

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

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

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.
- 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.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Best subset selection

- Fit all possible models $(2^p - 1)$ and select a single best model from according certain criteria.
- Possible criteria include adjusted R^2, cross-validated prediction error, C_p, AIC, or BIC.
- We consider the adjusted R^2 statistics

$$R^2_{adj} = 1 - \frac{SSE/(n-q-1)}{SST/(n-1)},$$

where q is the number of predictors in the model.

- Adjusted R^2 criterion: we pick the best model by maximizing the adjusted R^2 over all $2^p - 1$ models.
- R^2 is 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.

AIC Criterion

- The AIC statistics for a model is defined as

$$AIC = -2l(y) + 2(q+1) \overset{LM}{=} n\log(SSE/n) + 2(q+1),$$

  where $l(y)$ is log-likelihood of y and q is the number of predictors in the model.
- The first part of AIC statistic decreases as the number of predictors in the model q increases.
- The second part increases as q increases. This part is to penalize larger models.
- The AIC statistics is not necessary to decrease or increase as q increases.
- AIC criterion: pick the best model by minimizing AIC criterion over all models.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

BIC Criterion

- The BIC statistics for a model is defined as

$$BIC = -2l(y) + \log(n)(q+1) \overset{LM}{=} n\log(SSE/n) + \log(n)(q+1),$$

  where $l(y)$ is log-likelihood of y and q is the number of predictors in the model.
- Similar to AIC statistics, the BIC statistics adds the second part to penalize larger models.
- BIC criterion: pick the best model by minimizing BIC criterion over all models.
- The only difference between AIC and BIC is the coefficient for the second part.
- The BIC criterion can guarantee that we can pick all the important predictors as $n \longrightarrow \infty$, while the AIC criterion cannot.

Cross-Validation

- The idea of cross-validation (CV) criterion is to find a model which minimizes the prediction/testing error.
- 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 $\text{CV} = \sum_{i=1}^{n}(y_i - \hat{y}_{-i})^2$ statistics 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".
- So CV statistics is simplified to

$$\text{CV} = \sum_{i=1}^{n}\left(\frac{r_i}{1-h_{ii}}\right)^2,$$

where h_{ii} is the ii-th element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Mallow's C_p Statistic

- The C_p statistics is another statistic which penalizes larger model. In the original definition, p is the number of predictors in the model. Unfortunately, we use q to denote the number of predictors. In the following we use the notation C_q instead.
- The C_q statistics for a given model is defined as

$$C_q = \frac{SSE(q)}{SSE(p)/(n-p-1)} - (n - 2(q+1)).$$

- It can be shown that $C_q \approx q + 1$, if all the important predictors are in the model.
- C_q criterion: pick the model such that C_q is close to $q + 1$ and also q is small (we like simpler model).
- In linear model, under Gaussian error assumption C_p criterion is equivalent to AIC.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Backward Elimination

- 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 the single F-test. Other criteria, say, AIC, BIC, C_p, apply as well.
- Check whether the p-values for all the p predictors are smaller than α, called alpha to drop.
- If yes, stop the algorithm and all the p predictors are treated as important.
- If not, delete the least significant variable, i.e., the variable with the largest p-value and repeat checking.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Forward Selection

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

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

For each predictor, we calculate the p-value of the single F-test for the hypothesis $H_0 : \beta_1 = 0$. Other criteria, say, AIC, BIC, C_p, apply as well.

- Choose the most significant predictor, denoted by $x_{(1)}$ such that the p-value of the F-test statistic for the hypothesis $H_0 : \beta_1 = 0$ is smallest.
- If the p-value for the most significant predictor is larger than α (alpha to add). We stop and no predictor is needed.
- If not, the most significant predictor is added in the model and we repeat choosing.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Stepwise selection

- 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 re-assessed at later steps.
- Stepwise selection, 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.
- Backward stage: Do Backward Selection until stop.
- Continue until no predictor can be added and no predictor can be removed according to the specified α to enter and α to drop.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Summary and Remark

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