Model Selection: Criterion Based methods

ST552 Lecture 22

Charlotte Wickham

2019-03-01

Stepwise methods

Stepwise methods consider a single path through the models (and only ever one model of each size).

An alternative is to consider all possible models (a.k.a Best Subset), but we need a way to compare models (F-tests don't help for non-nested models).

General strategy of best subsets: calculate metric for all possible models. Choose model with the "best" value of the metric

Some common metrics, in the regression setting

$$\text{Akaike Information Criterion} = \text{AIC} = n \log\left(RSS/n\right) + 2p$$

$$\text{Bayesian Information Criterion} = \text{BIC} = n \log\left(RSS/n\right) + p \log n$$

$$\text{Mallow's Cp} = \frac{RSS}{\hat{\sigma^2}} + 2p - n$$

where $\hat{\sigma^2}$ is from the full model

$$\text{Adjusted R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$$

Small is good for AIC, BIC and Mallow's Cp.

Large is good for adjusted R^2.

Which is best? It depends. . .

AIC

Arises from considering how to estimate the distance of a candidate model from the true model.

In particular using the Kullback-Leibler information to measure distance

$$I(f,g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx$$

In general, can estimate the expected value, at the maximum likelihood estimate of θ with

$$-\log L(\hat{\theta}) + p + \text{constant}$$

Akaike multiplied this by 2. We can ignore any constants that are the same for a given dataset and assumed error distribution for regression variable selection. Care should be taken in other contexts.

BIC

Arises from trying to find the model with the highest posterior probability.

Mallow's Cp

Tries to estimate the mean square prediction error.

Adjusted R-squared

Find a model with the highest R^2, but R^2 always increases when you add a variable. Adjusted R^2 penalises for more variables.

Return to example in R

Limitations of best subsets methods (or criterion based methods as Faraway calls them)

1. p-values will generally overstate the importance of remaining predictors
2. Inclusion in the model doesn't correspond to important, and exlcusion doesn't correspond to unimportant.

Comments

These metrics are estimates, and like all estimates are subject to variability.

The ranking of models for one dataset might be different to another generated from the same data generating process.

There are some assymptotic results. Two common types:

- consistent for model selection: if you have enough data you will get the right model
- optimal for prediction: if you have enough data you will get the best predictions (in the sense of squared error)

Alternative estimate of model performance: use external test set, and estimate your desired metric directly. (The idea behind cross validation)