Model Selection

Frank Wood

December 10, 2009

Standard Linear Regression

Recipe

-
  - Identify the explanatory variables
  - Decide the functional forms in which the explanatory variables can enter the model
  - Decide which interactions should be in the model.
- Reduce explanatory variables (!?)
- Refine model

Trouble and Strife

Form any set of $p-1$ predictors, 2^{p-1} alternative linear regression models can be constructed

- Consider all binary vectors of length $p-1$

Search in that space is exponentially difficult.

Greedy strategies are typically utilized.

Is this the only way?

Selecting between models

In order to select between models some score must be given to each model.

The likelihood of the data under each model is not sufficient because, as we have seen, the likelihood of the data can always be improved by adding more parameters until the model effectively memorizes the data.

Accordingly some penalty that is a function of the complexity of the model must be included in the selection procedure.

There are four choices for how to do this

1. Explicit penalization of the number of parameters in the model (AIC, BIC, etc.)
2. Implicit penalization through cross validation
3. Bayesian regularization / Occam's razor.
4. Use a fixed model of unbounded complexity (Bayesian nonparametrics).

Penalized Likelihood

The Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) (also called the Schwarz criterion) are two criteria that penalize model complexity.

In the linear regression setting

$$\begin{aligned} AIC_p &= n \ln SSE_p - l \ln n + 2p \\ BIC_p &= n \ln SSE_p - n \ln n + (\ln n)p \end{aligned}$$

Roughly you can think of these two criteria as penalizing models with many parameters (p in the case of linear regression).

[1] For a good derivation of the BIC see http://myweb.uiowa.edu/cavaaugh/ms_lec_6_ho.pdf

Cross Validation

A way of doing model selection that implicitely penalizes models of high complexity is cross-validation.

If you fit a model with a large number of parameters to a subset of the data then predict the rest of the data using the model you just fit, then

- The average scores of held-out data over different *folds* can be used to compare models.
- If the held-out data is consistently (over the various folds) well explained by the model then one could conclude that the model is a good model.
- If one model performs better on average when predicting held-out data then there is reason to prefer that model.
- Overly complex models will not generalize well and will therefor not be selected.

$PRESS_p$ or Leave-One-Out Cross Validation

The $PRESS_p$ or "prediction sum of squares" measures how well a subset model can predict the observed responses Y_i.

Let $\hat{Y}_i(i)$ be the fiitted value when i is being predicted from a model in which (i) was left out during training.

The $PRESS_p$ criterion is then given by summing over all n cases

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2$$

$PRESS_p$ values can be calculated without doing n separate regression runs.

$PRESS_p$ or Leave-One-Out Cross Validation

If we let d_i be the *deleted residual* for the i^{th} case

$$d_i = Y_i - \hat{Y}_{i(i)}$$

then we can rewrite

$$d_i = \frac{e_i}{1 - h_{ii}}$$

where e_i is the ordinary residual for the i^{th} case and h_{ii} is the i^{th} diagonal element in the hat matrix.

We can obtain the h_{ii} diagonal element of the hat matrix directly from

$$h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$$

$PRESS_p$ or Leave-One-Out Cross Validation

$PRESS_p$ is useful for choosing between models. Consider

- Take two models, one $\mathcal{M}_p$ with p dimensional input and one $\mathcal{M}_{p-1}$ with $p-1$ dimensional input
- Calculate the $PRESS_p$ criterion for $\mathcal{M}_p$ and $\mathcal{M}_{p-1}$
- Whichever model has the lowest $PRESS_p$ should be preferred.
- Why?

Unfortunately the $PRESS_p$ criteria can't tell us which variables to include. For that general search is still required.

More general cross-validation procedures can be designed.

Note that the $PRESS_p$ criterion is very similar to log-likelihood.

Detecting Outliers Via Studentized Deleted Residuals

The *studentized deleted residual*, denoted by t_i is

$$t_i = \frac{d_i}{s\{d_i\}}$$

where

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1 - h_{ii}} \qquad \frac{d_i}{s\{d_i\}} \sim t(n - p - 1)$$

Fortunately, again, t_i can be calculated without fitting n different regression models. It can be shown that

$$t_i = e_i \left[\frac{n - p - 1}{SSE(1 - h_{ii} - e_i^2)} \right]^{1/2}$$

These t_i's can be used to formally test (e.g. using a Bonferroni test procedure) whether the largest absolute studentized deleted residual is an outlier.

Stepwise Regression Methods

A (greedy) procedure for identifying variables to include in the regression model is as follows. Repeat until finished:

1. Fit a simple linear regression model for each of the P-1 X variables considered for inclusion. For each compute the t^* statistics for testing whether or not the slope is zero

$$t_k^* = \frac{b_k}{s\{b_k\}}^2$$

2. Pick the largest out of the $P-1$ t_k^*'s (in the first step $k=1$) and include the corresponding X variable in the regression model if t_k^* exceeds some arbitrary threshold.
3. If the number of X variables included in the regression model is greater than one, check to see if the model would be improved by dropping variables (using the t-test and a threshold again).

[2] Remember b_k is the estimate for β_k and $s\{b_k\}$ is the estimator sample standard deviation.

Stepwise Regression Methods

Big question: Does this ensure the "best" possible model? Either in the "is this procedure guaranteed to pick the best possible linear regression model" or in any other sense?

There is a tension between including variables (why not include everything?) and needing to reliably estimate many parameters. Here sharp philosophical differences emerge (partially driven by the needs of the user / application)

1. Try to identify which elements are linearly related to the output.
2. Include everything and regularize.

Finding the Best Model, Case 1

When there are many parameters and few observations this is known as the big p little n problem.

One might actually to know (the inference goal) which inputs are linearly related to the output.

It is tempting but *dangerous* and *wrong* to conclude that if by formal testing procedures you find that a particular input feature is not linearly related to the output that there is no relationship between the variables.

A converse is also potentially *dangerous* and *wrong*, namely, if you find that a particular feature has a "statistically significant" linear effect on the output you cannot necessarily conclude causality. Carefully controlled experiments are required to establish probable causality.

Finding the Best Model, Case 2

Philosophically it makes sense to include all possible features in a regression model. Why not?

Unfortunately, in the big p small n setting model estimation is usually difficult. In linear regression the sample covariance matrix is low rank and not-invertible so the small n big p problem is degenerate.

Regularization can be employed, but it is difficult to interpret and assign "physical" meaning to the resulting regression coefficients.

This may not matter if prediction is the goal, rather than describing or examining the relationship between the predictors and the output.

Finding the Best Model, Alternative

It is possible to include all the variables in the model and automatically learn which variables should be included.

Techniques for doing this go by different names (LASSO, L1 penalized regression, etc.).

They use a different regularizer (L1 instead of L2) term that encourages sparsity (i.e. zero parameters).

Fitting such a model becomes significantly more difficult and requires using a constrained optimization solver.