Predictive Analytics Lecture 6

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

February 21 & 22, 2017

Modeling Framework Refresher

Recall the general regression model:

$$Y = f(x_1, \ldots, x_p) + \mathcal{E}$$

A couple lectures ago, we made the parametric assumption that:

$$Y = s(x_1, \ldots, x_p; \theta_1, \ldots, \theta_\ell) + \tilde{\mathcal{E}}$$

where the $\tilde{\mathcal{E}}$ term now includes the previous $\mathcal{E}$ plus $f - s$, the misspecification error. The parametric model s we employed was the linear model and the θ's we called β's:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \tilde{\mathcal{E}}$$

Last lecture, we started adding interactions and polynomials (as well as other transformations e.g. log which we did not cover). This was a means of "expanding" the feature set "visible" to the model using "derived" features:

$\{x_1, \ldots, x_p\} \Rightarrow \{x'_1, \ldots, x'_{p'}\}$ where $p' > p$ and maybe much, much greater.

“Non-parametric” Linear Regression

Once we expand this feature set, we can now fit a larger linear model:

$$Y = \beta_0 + \beta_1 x_1' + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \beta_p x_{p'}' + \tilde{\mathcal{E}}$$

Given more degrees of freedom with this expanded feature set allows the linear model to fit more complicated real-world functions. This is essentially a means of doing non-parametric parametric modeling (it’s oxymoronic). It’s technically parametric but conceptually it’s non-parametric since we don’t have our parametric benefits: parsimony, inference nor interpretation. Hopefully $\tilde{\mathcal{E}}$ will be close to $\mathcal{E}$, the irreducible noise.

Back to our problem... we can curb overfitting by ... using 3-way split oos validation but we need to select good models... how to do so? One approach is termed **subset selection methods**.

Stepwise Regression

First we expand the feature set from $\{x_1, \ldots, x_p\} \Rightarrow \left\{x'_1, \ldots, x'_{p'}\right\}$. Then we attempt to find the "best" model consisting of a subset of these features. However there are $2^{p'}$ possible models. For $p' = 20$ that's about 1,000,000. So we try to find a model *close* to the optimal using a "heuristic" (a rule of thumb that seems to generally be useful).

That heuristic is called **stepwise** model construction. We begin with **forward stepwise** model construction:

1. Find the "best" feature from the list of expanded features.
2. Find the "next best" feature from the remaining expanded features.
3. Repeat step 2 until you believe you are overfitting.

Estimating Overfitting (again)

If you choose the feature to give you the best in-sample R^2, you will eventually take all the features (until $n = p + 1$) and you will get $R^2 = 100\%$. We need a metric to tell us when we may be overfitting and halt at that moment. Here are a few:

1. oos RMSE (keep a holdout set and quit when this starts increasing)
2. Only include a variable if its t stat (or partial F stat) is significant
3. Use $AICc$.

$$-AIC = 2\ell\left(\hat{\beta}; \boldsymbol{y}, \boldsymbol{x}\right) - 2p$$

The first component (the log-likelihood) represents in-sample fit. $\ell()$ is like R^2 though... as the fit gets closer to the points, the likelihood goes to 1 (and the log likelihood goes to 0). The $2p$ term is a reality check. If you have more features, you are going to overfit. So each additional feature must be justified in terms of the increase in log-likelihood. Thus, good models maximize $-AIC$ (i.e. minimize AIC).

AICc for linear models

For linear regression under OLS, we can calculate the log-likelihood explicitly (we approximately did this in Lecture 2) to obtain:

$$AIC = n \ln \left(RMSE^2 \right) + 2p$$

So once again, we want this to be small. If we decrease *RMSE* by adding a feature, it needs to counteract an increase of 2 by $p \to p + 1$. If it can't, we're probably overfitting. *AIC* works well with large sample sizes. For small sample sizes, we use a corrected version *AICc* defined as:

$$AICc = AIC + \frac{2p(p+1)}{n - p - 1}$$

Needless to say, this is all approximate since we are assuming OLS and a whole bunch of other things (beyond scope of course). Note: there are also BIC and Mallow's C_p which are similar metrics, but we will not cover them. (JMP demo for white wine dataset and telecom set.)

More About Stepwise Linear Regression

Backward selection begins with all features and then deletes one for each step until no more can justifiably be cut out. Backward selection has a major weakness: it cannot be run on dataframes where the extended feature set is more than the number of rows (only forward or forward with mixed works there). **Mixed selection** begins with either none or all and then looks for both good additions and good subtractions.

Simple case where stepwise doesn't work? How about three features where x_1 is most correlated but x_2 and x_3 together are the best model but there is high collinearity between x_2 and x_3? What happens? Forward: the model enters x_1 and then x_2 but it doesn't see x_3 as a worthy addition. Backward: the model can nuke x_2 or x_3 since its p-value or F test is poor.

This is not the only way to fit a flexible model and hedge against overfitting. We will do more such models. But first, we will talk about "missing data" as it's more relevant to the project which is due soon.