Class 7 Review Notes

AI & Machine Learning

Fall 2023 - Laurie Ye

7 Model Building Introduction

7.1 introduction

Model building (variable selection) involves:

1. Determining a set of candidate X-variables (independent variables) believed to have potential to "explain" the dependent Y variable.
2. Selecting a subset of these variables that either best explains or explains very well the dependent Y variable.

7.2 First Step - Preprocessing

Model building (variable selection) involves:

1. There is usually important multicolinearity.
2. Transformations to symmetry should be done as this tends to linearize and simplify the relationships.
3. Categorical variables are generally the most common type of variable. If there are k categories, then k-1 indicator (dummy) variables are needed to encode them. High k means high dimension.
4. Polynomial and interaction terms can capture non-linearity, but they often behave very badly outside of the data. - handling non-linearity via transformation when possible is generally better.

Note 1. The conclusion is that you should, to the extent possible, understand the variables (features) and think hard about what form they should take.

7.3 Variable Selection

There are generally three approaches taken:

1. Forward stepwise: Start with a small model and add in the next most important variable one at a time.

2. Backwards stepwise: Start with all variables in the model and remove the next least important variable one at a time.

3. Best subsets or all subsets: Search through all possible subsets of the variables for the best or top few sets.

Comments

- Forward stepwise selection can allow backwards steps that remove variables entered in the model that become unimportant.

- Backwards stepwise selection can allow forward steps that consider entering variables previously deleted that have become important.

- In general, if you put in an interaction term, you should put in the main effects as well (even if not significant).

- In general, if you put in a higher order polynomial term, the lower order terms should be put in the model even if they are not significant. So if X^2 is in the model, then X should be in the model as well.

- In ML problems with a large number of X-variables, backwards stepwise may not be feasible since you basically cannot fit a model with everything in.

- Your criterion for whether or not to include or drop variable SHOULD take into account the number of variables (model complexity). Usually we will use AIC or AICc.

- In ML, we will use out-of-sample cross-validation to select the final model.

- But, the more models we "throw" at the validation sample, the more likely it is that there will be some overfitting going on in the selection of the best model. A reasonable strategy is to use intra-sample (within the training sample) variable selection measures like AIC or AICc to select a few good models and then use cross validation to choose from among these.

  Note 2. In ML other approaches can be used (such as regularization) that tend to leave everything in the model, but penalize model complexity.

Sanity Check

- Are the coefficients the correct sign?

- Does the model make sense?

- Is a prediction you are trying to make within the X-data or outside (i.e., an extrapolation). It can be hard to know in high dimensions.

7.4 AIC and AICc

- AIC = Akaike Information Criterion. This is an asymptotic approximation to a "distance" measure between the data and the model.
- AICc is a better approximation in small samples.
- For both AIC and AICc, smaller is better.
- You cannot interpret the magnitude of the AIC or AICc.
- To be valid AIC and AICc must be used on the same data to compare models with different sets of X-variables.
- The dependent variable must be the same and there must be the same number of observations.

Note 3. The magnitude is not interpretable. But the smallest AIC/AICc is the best model among the set of models considered.

7.5 Reference

Goizueta Business School-Emory University: Professor George S. Easton