What are some useful guidelines for GBM parameters?

What are some useful guidelines for testing parameters (i.e. interaction depth, minchild, sample rate, etc.) using GBM?

Let's say I have 70-100 features, a population of 200,000 and I intend to test interaction depth of 3 and 4. Clearly I need to do some testing to see what combination of parameters holds up best out-of-sample. Any suggestions on how to approach this test design?

r | cart | hypothesis-testing | package | gbm

asked Apr 3 '12 at 3:27

Quant Guy
908 ● 10 ■ 27

1 Answer

The caret package can help you optimize the parameter choice for your problem. The caretTrain vignette shows how to tune the gbm parameters using 10-fold repeated cross-validation - other optimization approaches are available it can all run in parallel using the foreach package. Use `vignette("caretTrain", package="caret")` to read the document.

The package supports tuning `shrinkage`, `n.trees`, and `interaction.depth` parameters for the gbm model, though you can add your own.

For heuristics, this is my initial approach:

`shrinkage`: As small as you have time for (the gbm manual has more on this, but in general you can nver go wrong with a smaller value). Your data set is small so I'd probably start with 1e-3

`n.trees`: I usually grow an initial model adding more and more trees until `gbm.perf` says I have enough (actually, typically to 1.2 times that value) and then use that as a guide for further analysis.

`interaction.depth`: you already have an idea about this. Try smaller values as well. Maximum value is floor(sqrt(NCOL(data)).

`n.minobsinnode`: I find it really important to tune this variable. You don't want it so small that the algorithm finds too many spurious features.

answered Apr 3 '12 at 11:03

Allan Engelhardt
311 ● 2 ■ 2