Does Breiman's random forest use information gain or Gini index?

I would like to know if Breiman's random forest (random forest in R randomForest package) uses as a splitting criterion (criterion for attribute selection) information gain or Gini index? I tried to find it out on http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm and in documentation for randomForest package in R. But the only thing I found is that Gini index can be used for variable importance computing.

r random-forest entropy gini

edited Apr 4 '15 at 16:36 Nick Cox 27.6k 3 54 82

asked Apr 4 '15 at 16:17 somebody 41 2

I also wonder if trees of random forest in randomForest package are binary or not. – somebody Apr 4 '15 at 16:51

1 Answer

The randomForest package in R by A. Liaw is a port of the original code being a mix of c-code(translated) some remaining fortran code and R wrapper code. To decide the overall best split across break points and across mtry variables, the code uses a scoring function similar to gini-gain:

$$GiniGain(N,X) = Gini(N) - \frac{|N_1|}{|N|}Gini(N_1) - \frac{|N_2|}{|N|}Gini(N_2)$$

Where X is a given feature, N is the node on which the split is to be made, and N_1 and N_2 are the two child nodes created by splitting N. $|.|$ is the number of elements in a node.

And $Gini(N) = 1 - \sum_{k=1}^{K} p_k^2$, where K is the number of categories in the node

But the applied scoring function is not the exactly same, but instead a equivalent more computational efficient version. $Gini(N)$ and |N| are constant for all compared splits and thus omitted.

Also lets inspect the part if the sum of squared prevalence in a node(1) is computed as $\frac{|N_2|}{|N|}Gini(N_2) \propto |N_2|Gini(N_2) = |N_2|(1 - \sum_{k=1}^{K} p_k^2) = |N_2| \sum \frac{nclass_{2,k}^2}{|N_2|^2}$

where $nclass_{1,k}$ is the class count of target-class k in daughter node 1. Notice $|N_2|$ is placed both in nominator and denominator.

removing the trivial constant $1-$ from equation such that best split decision is to maximize nodes size weighted sum of squared class prevalence...

score= $|N_1| \sum_{k=1}^{K} p_{1,k}^2 + |N_2| \sum_{k=1}^{K} p_{2,k}^2 = |N_1| \sum_{k=1}^{K} \frac{nclass_{1,k}^2}{|N_1|^2} + |N_2| \sum_{k=1}^{K} \frac{nclass_{2,k}^2}{|N_2|^2}$
$= \sum_{k=1}^{K} \frac{nclass_{2,k}^2}{1} |N_1|^{-1} + \sum_{k=1}^{K} \frac{nclass_{2,k}^2}{1} |N_1|^{-2}$
$= nominator_1/denominator_1 + nominator_2/denominator_2$

The implementation also allows for classwise up/down weighting of samples. Also very important when the implementation update this modified gini-gain, moving a single sample from one node to the other is very efficient. The sample can be substracted from nominators/denominators of one node and added to the others. I wrote a prototype-RF some months ago, ignorantly recomputing from scratch gini-gain for every break-point and that was slower :)

If several splits scores are best, a random winner is picked.

This answer was based on inspecting source file ***"randomForest.x.x.tar.gz/src/classTree.c"*** line 209-250

edited Aug 14 '15 at 18:33

answered Aug 14 '15 at 14:00 Soren Havelund Welling 2,921 6 19

Add Another Answer