2,871 5 17 review help

number of trees that were built without minority class?

Lets assume that my random forest has 500 trees. My data is imbalance with 90% of class A and 10% of class B. I am wonder if there is any way to calculate roughly the number of trees that are built with only samples from class A.

Thanks

random-forest cart

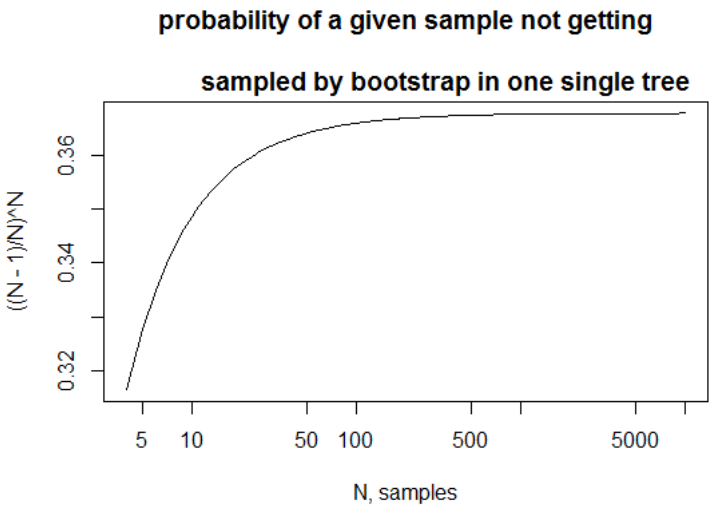
asked Feb 4 at 20:31

rudky martin
16 2

1 Answer

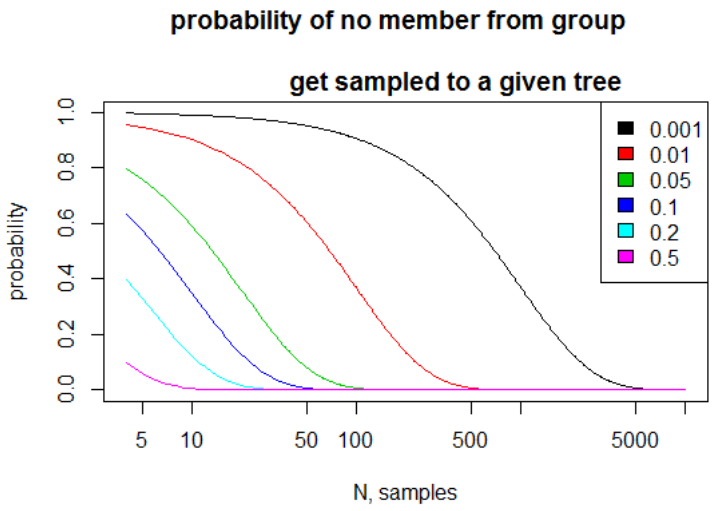
You can also have the exact answer. Besides sizes of groups it depends on size of training set (*N*). First I compute the probability that any given sample *x* is not selected for a tree. For a given tree *N* samples are drawn and the probability *(one draw not x)* $= \frac{N-1}{N}$ and the probability of *(all draws not x) = (one draw not x)* ^{N}.

```
N=unique(ceiling(10^((seq(.5,4,len=500)))))
plot(N,((N-1)/N)^N,log="x",type="l",xlab="N, samples",
     main="probability of a given sample not getting\n
           sampled by bootstrap in one single tree")
PnotSample = (((N-1)/N)^N)
```



Next I define 6 scenarios where the small group *X* make out 0.1%, 1%, 5%, 10%, 20% and 50% for any number of training set size. *(all draws not group X) = (all draws not x)* ^{NgroupRatio}

```
#probability of group not getting sampled to a tree
groupRatio=c(0.001,.01,.05,.1,.2,.5)
for(i in 1:length(groupRatio)) {
  PnotGroup = PnotSample^(N*groupRatio[i])
  if(i==1) {
    plot(N,PnotGroup,log="x",type="l",col=i,
         xlab="N, samples",ylab="probability",
         main="probability of no member from group  \n
               get sampled to a given tree")
  } else {
    points(N,PnotGroup,type="l",col=i)
  }
}
legend("topright",legend=groupRatio,fill=1:length(groupRatio))
```



So in your case (blue line), not selecting any sample from minor group would be quite rare if you have more than 50 samples in your training set. The expected number of trees not having any from minor group is simply the probability for a single tree multiplied with number of trees in forest. Anyhow, there could be a number of good reasons to modify these odds, you can read about why and how in this answer.

edited Feb 7 at 14:47

answered Feb 7 at 14:34

Soren Havelund Welling
2,871 5 17

Add Another Answer