Boosting

Vinoth Aryan Nagabosshanam

March 25, 2017

Boosting

Boosting builds lots of smaller trees. Unlike random forests, each new tree in boosting tries to patch up the deficiencies of the current ensemble.

```
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 3.3.3
```

```
## Loading required package: survival
```

```
## Loading required package: lattice
```

```
## Loading required package: splines
```

```
## Loading required package: parallel
```

```
## Loaded gbm 2.1.3
```

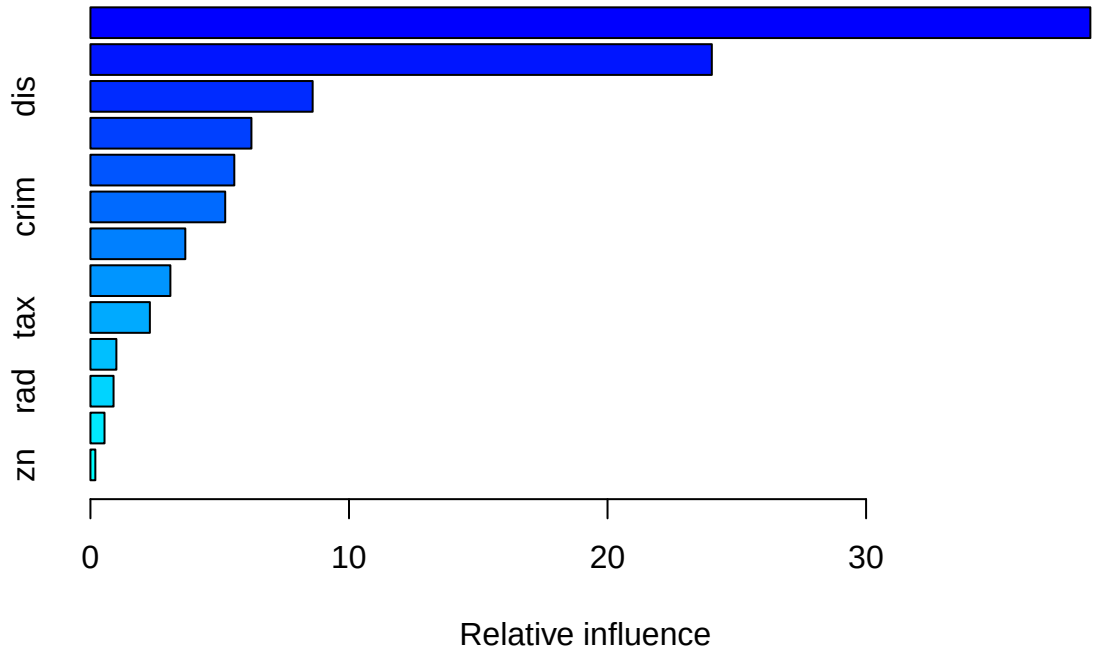
```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.3.3
```

```
library(MASS)
```

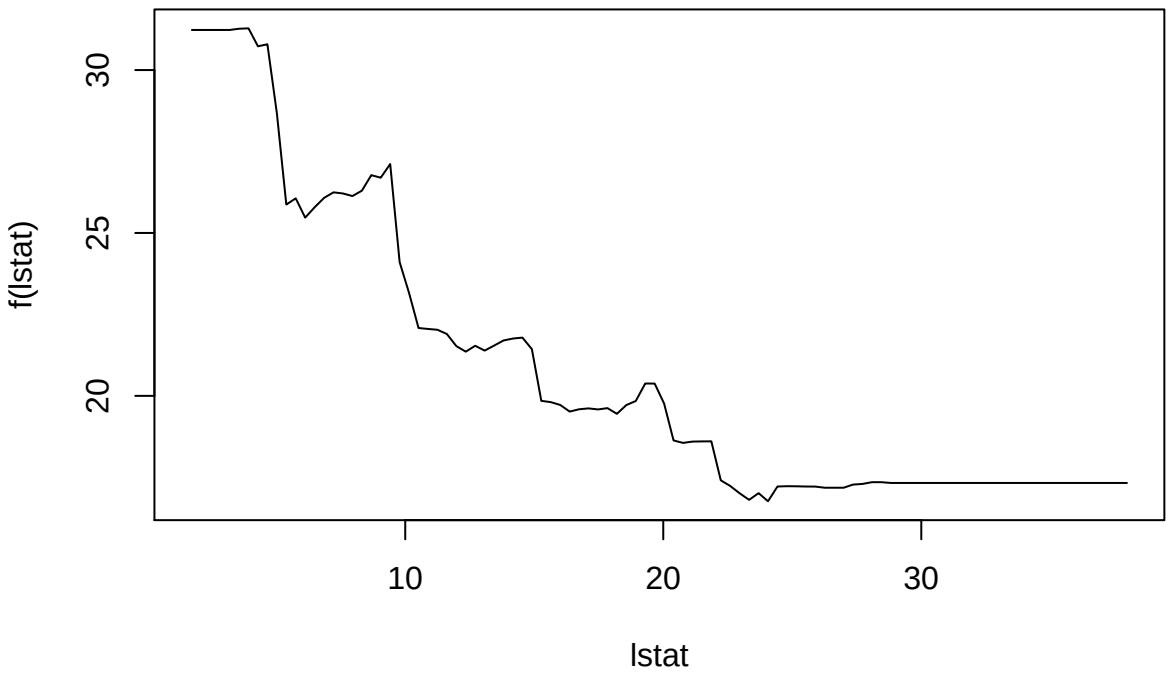
```
## Warning: package 'MASS' was built under R version 3.3.3
```

```
#View(Boston)
train=sample(1:nrow(Boston),300)
boost.boston=gbm(medv~.,data=Boston[train,],distribution="gaussian",n.trees=10000,shrinkage=0.01,intera
summary(boost.boston)
```

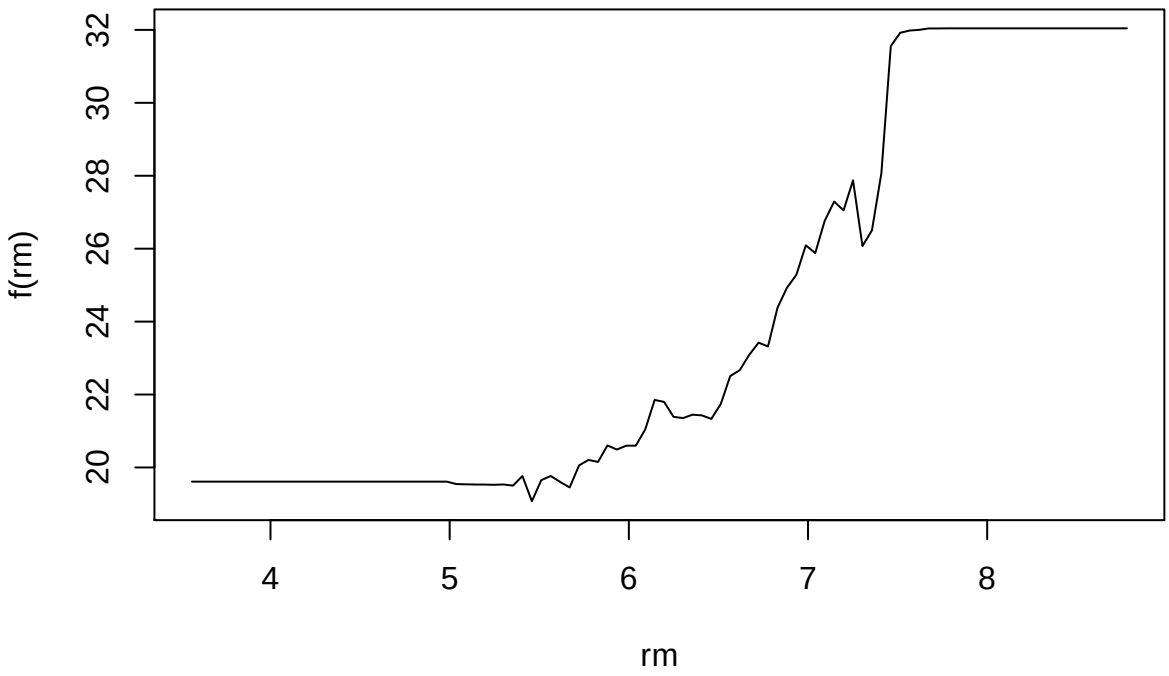



```
##             var    rel.inf
## lstat     lstat 38.6743820
## rm           rm 24.0341064
## dis         dis  8.5945754
## nox         nox  6.2267140
## age         age  5.5646177
## crim       crim  5.2125267
## black     black  3.6692445
## ptratio ptratio  3.0925433
## tax         tax  2.2994252
## indus     indus  1.0031019
## rad         rad  0.8944638
## chas       chas  0.5448754
## zn           zn  0.1894238
```

```
plot(boost.boston,i="lstat")
```




```
plot(boost.boston,i="rm")
```

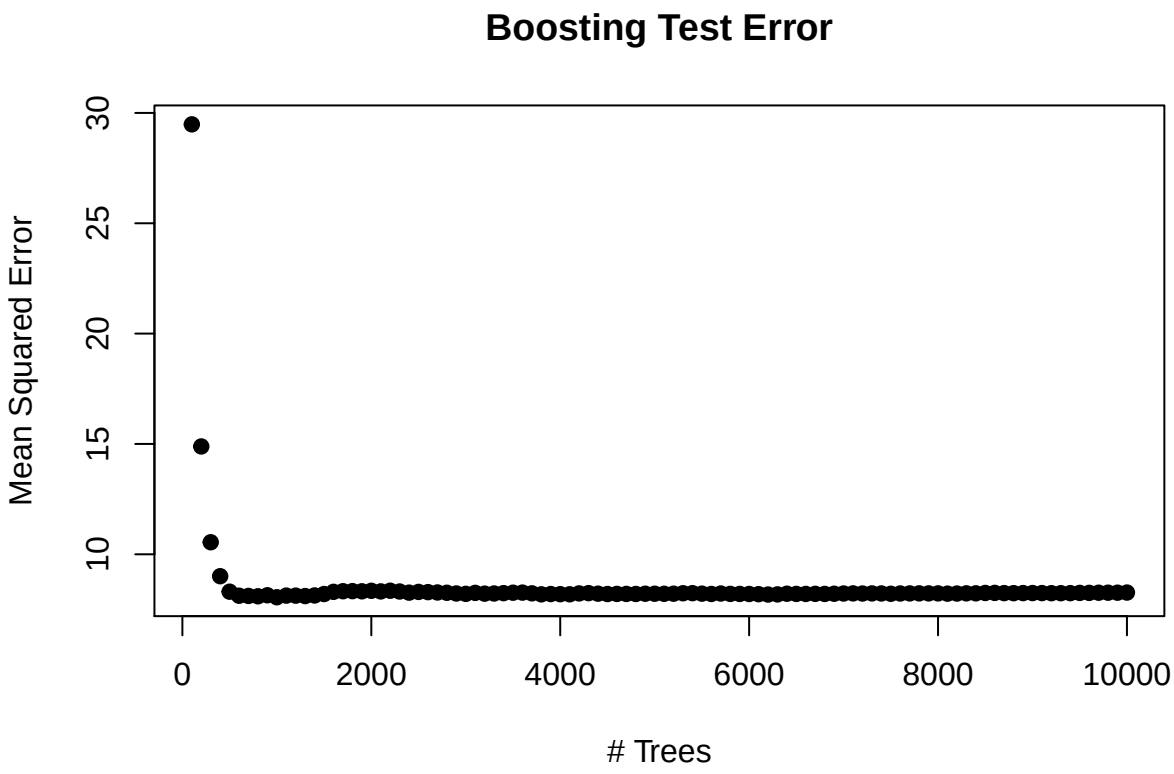


Lets make a prediction on the test set. With boosting, the number of trees is a tuning parameter, and if we have too many we can overfit. So we should use cross-validation to select the number of trees. We will leave this as an exercise. Instead, we will compute the test error as a function of the number of trees, and make a plot.

```
test.err=double(13)
n.trees=seq(from=100,to=10000,by=100)
predmat=predict(boost.boston,newdata=Boston[-train,],n.trees=n.trees)
dim(predmat)
```

```
## [1] 206 100
```

```
berr=with(Boston[-train,],apply( (predmat-medv)^2,2,mean))
plot(n.trees,berr,pch=19,ylab="Mean Squared Error", xlab="# Trees",main="Boosting Test Error")
abline(h=min(test.err),col="red")
```


Boosting Test Error
Mean Squared Error
10
15
20
25
30
0
2000
4000
6000
8000
10000
Trees