Classification and Regression Trees with ada and gbm

Samuel Croker

August 23, 2013

Stochastic Boosting

- Supervised learning
- Algorigthm - Schapire 1990 (AdaBoost 1996)
- Ensemble of weak learners
- Works well for categorical features

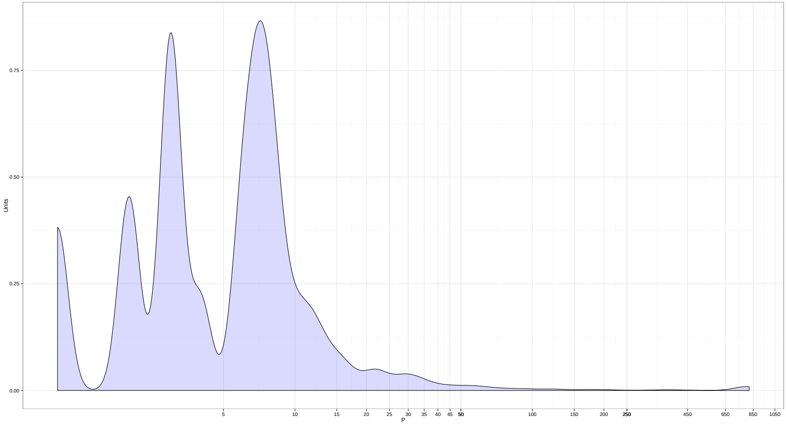
R Packages for Stochastic Boosting

- `ada` - Discrete, simple implementation
- `gbm` - Generalized boosting, regression
- `mboost` - Generalized boosting, regression
- Other Suggestions?

Source Data

```
> str(stwX)
'data.frame': 10691 obs. of  13 variables:
 $ XD1: int  4 6 3 6 2 6 4 4 2 3 ...
 $ YD : chr  "O" "O" "O" "O" ...
 $ XD2: Factor w/ 3 levels "...",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ XD3: Factor w/ 2 levels "","UNK": 2 2 2 2 2 2 2 2 2 2 ...
 $ XD4: Factor w/ 41 levels "AF",..: 13 13 11 15 13 11 15 9 9 11 ...
 $ XC1: num  2296 295 3298 136 1692 ...
 $ XC2: int  7 9 7 5 5 8 8 9 39 5 ...
 $ Y  : int  8 8 7 11 18 6 6 6 17 14 ...
 $ XCN: int  1 1 1 1 1 1 1 1 1 1 ...
 $ XD5: Factor w/ 18 levels "...",..: 5 9 17 10 10 7 17 17 17 18 ...
 $ XD6: Factor w/ 7 levels ".",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ XD7: Factor w/ 3 levels "I","O","Other": 2 2 2 2 2 2 2 2 2 2 ...
 $ rnd: num  0.485 0.48 0.987 0.185 0.158 ...
```

Data Density



Data Preperation

```
stwX <- transform(stwX,YD=ifelse(Y<=5,’E’,’O’))
stwX <- data.frame(stwX,rnd=runif(length(stwX[,1])))

est <- subset(stwX,rnd<0.8)[,c(1,2,4,5,6,7,10:12)]
val <- subset(stwX,rnd >=0.8)[,c(1,2,4,5,6,7,10:12)]

n <- length(est[,1])
train<-sample(1:n,floor(.7*n),FALSE)
test<-setdiff(1:n,train)
```

ada Call

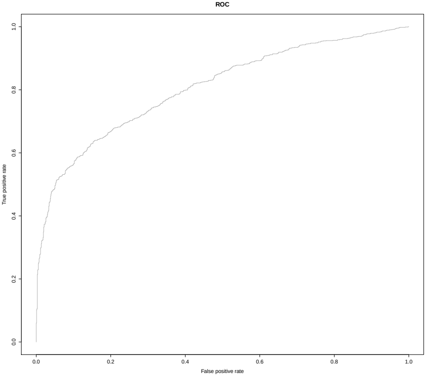
```
bt.fit <- ada(YD ~ .
              , data = est[train,],
              iter=500
              ,nu=.1
              ,type='discrete')

bt.fit <- addtest(bt.fit,test.x=est[test,-2]
                         ,test.y=est[test,2])
```

Evaluating Training Step

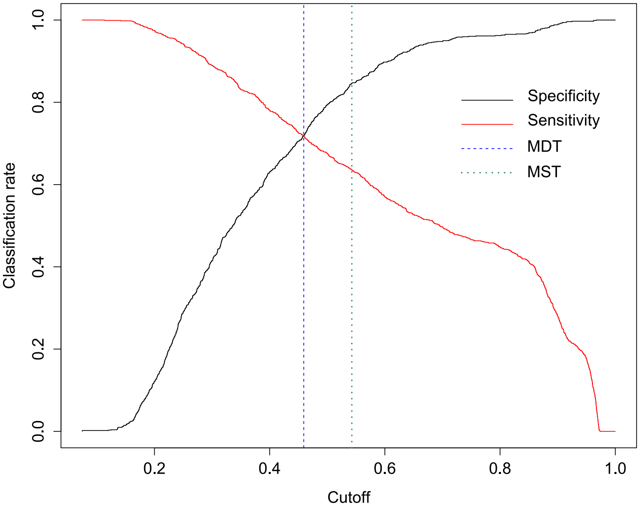
```
> summary(bt.fit)
Call:
ada(YD ~ ., data = est[train, ]
            , iter = 500, nu = 0.1, type = "discrete")
Loss: exponential Method: discrete   Iteration: 500
Training Results
Accuracy: 0.737 Kappa: 0.477
Testing Results
Accuracy: 0.707 Kappa: 0.415
```

Identifying Optimal Cutoff I

```
> PredBTlook<- data.frame(obs=val$YD,predict=pred1$class)

> stcPred(PredBT$predict,PredBT$obs)
[1] "(MDT,MST) = ( 0.459209308105611 , 0.54264704394395 )"
```



Identifying Optimal Cutoff II



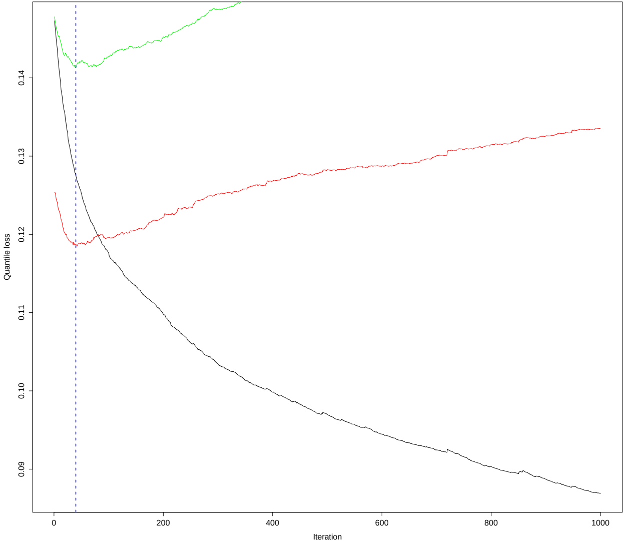
Accuracy, Sensitivity and Specificity

```
> precision(.45,PredBT$predict,PredBT$obs)
       obs
         <=5  >5
 P <=5   809 307
    >5   305 731
[[1]]


[[2]]
specificity sensitivity
  0.7055985   0.7249104

[[3]]
<= 5 Predictive         >5 Predictive
Accuracy                Accuracy
      0.7042389             0.7262118

[1] "BT Validation Area Under the Curve: 0.806690763343988"
```

gbm Call

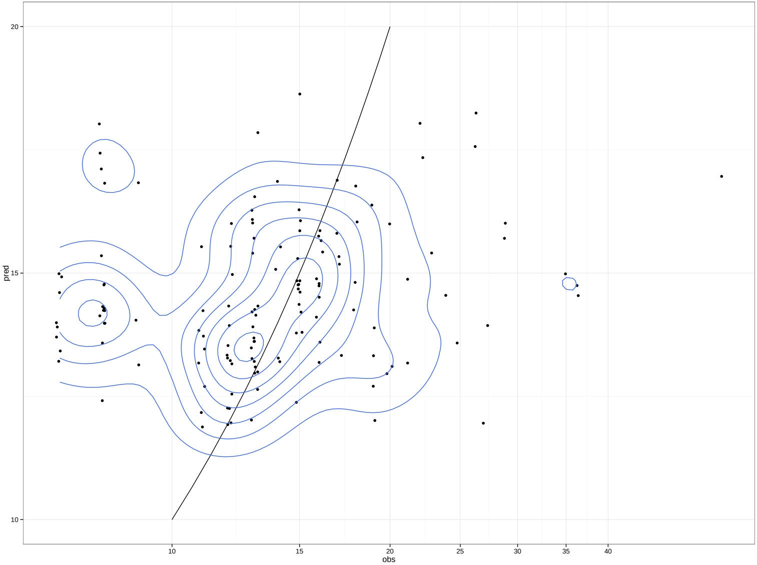
```
est <- subset(stwA,rnd < 0.8)[,c(1,4:8,10:12)]
val <- subset(stwA,rnd >=0.8)[,c(1,4:8,10:12)]

gbm.fit <- gbm(log(Y) ~ ., data=est,
    distribution=list(name=’quantile’,alpha=0.5),
    n.trees=1000,
    shrinkage=.05,
    interaction.depth=5,
    bag.fraction=.5,
    train.fraction=.5,
    cv.folds=5,
    keep.data=T,
    verbose=F
)
```

Fit Diagnostics



```
> best.iter <- gbm.perf(gbm.fit,method="cv")
> print(best.iter)
[1] 40

> summary(gbm.fit,n.trees=best.iter)
    var   rel.inf
XD5 XD5 34.994135
XD4 XD4 24.465712
XD1 XD1 21.217152
XC2 XC2  9.733716
XC1 XC1  7.938009
XD3 XD3  1.651275
XD6 XD6  0.000000
XD7 XD7  0.000000
```

Observed vs Predicted



```
ggplot(data=pre) +
  geom_jitter(aes(x=obs,y=pred)) +
  geom_line(data=z,aes(x=b,y=b)) +
  geom_density2d(aes(x=obs,y=pred)) +
  scale_x_continuous(trans=’log’,
          breaks=c(seq(0,40,5),75,100,150,200)) +
  scale_y_continuous(breaks=c(seq(0,20,5))) +
  theme_bw()
```

Suggested Reading

ECOL/BIOL 563 Statistical Methods in Ecology
http://www.unc.edu/courses/2010fall/ecol/563/001/

ada: An R Package for Stochastic Boosting; Culp, Johnson, Michailidis http://www.stat.wvu.edu/~mculp/math/ada/ada_manual.pdf

Generalized Boosted Models: A guide to the gbm package,Ridgeway
http://cran.open-source-solution.org/web/packages/gbm/vignettes/gbm.pdf

Visualizing Classifier Performance in R http://rocr.bioinf.mpi-sb.mpg.de/