

Machine Learning – IDM Sect 4.4 - 4.6

-Let's try crossvalidation with the sonar classification tree

```
train<-read.csv("sonar_train.csv",header=FALSE)
nxval <- 10
out <- matrix(nrow = nxval, ncol = 2)
l <- seq(from = 1, to = nrow(train))

for(idepth in seq(from = 1, to = 10)){
  trainErr <- 0.0
  testErr <- 0.0
  for(ixval in seq(from = 1, to = nxval)){
    lout <- which(l%%nxval == ixval%%nxval)
    trainIn <- train[-lout,]
    trainOut <- train[lout,]
    yin <- as.factor(trainIn[,61])
    yout <- as.factor(trainOut[,61])
    xin <- trainIn[,1:60]
    xout <- trainOut[,1:60]

    fit <- rpart(yin~,xin,control=rpart.control(maxdepth=idepth))
    trainErr <- trainErr + (1-sum(yin==predict(fit,xin,type = "class"))/length(yin))
    testErr <- testErr + (1-sum(yout==predict(fit,xout,type="class"))/length(yout))
  }
  out[idepth,1] <- trainErr/nxval
  out[idepth,2] <- testErr/nxval
}
```

Machine Learning – IDM Sect 4.4 - 4.6

- A tree can overfit SO CAN ALL OTHER ML METHODS
- How can we estimate the degree of underfit or overfit??
- Holdout Method



- K-fold cross validation



- Hold out 1st set, train on 2-k, then hold out 2 and train on 1 + 3-k etc.
- Calculate average error on training set and average error on test set.

Machine Learning – IDM Sect 4.4 - 4.6

