# 1 Random Forests

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

The algorithm for inducing a random forest was developed by Leo Breiman[1] and Adele Cutler,[2] and "Random Forests" is their trademark. The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995.

The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho[3][4] and Amit and Geman[5] in order to construct a collection of decision trees with controlled variance.

# 2 RandomForest with R

```
library(randomForest)

# download Titanic Survivors data
data <- read.table("http://math.ucdenver.edu/RTutorial/titanic.txt", h=T, sep="\t")
# make survived into a yes/no
data$Survived <- as.factor(ifelse(data$Survived==1, "yes", "no"))

# split into a training and test set
idx <- runif(nrow(data)) <= .75
data.train <- data[idx,]
data.test <- data[-idx,]
```

Train a random forest

```
rf <- randomForest(Survived ~ PClass + Age + Sex,
             data=data.train, importance=TRUE, na.action=na.omit)
```

How important is each variable in the model ?

```
imp <- importance(rf)
o <- order(imp[,3], decreasing=T)
imp[o,]
#             no      yes MeanDecreaseAccuracy MeanDecreaseGini
#Sex    51.49855 53.30255             55.13458         63.46861
#PClass 25.48715 24.12522             28.43298         22.31789
#Age    20.08571 14.07954             24.64607         19.57423
```

Display the confusion matrix

```
# confusion matrix [[True Neg, False Pos], [False Neg, True Pos]]
table(data.test$Survived, predict(rf, data.test),
  dnn=list("actual", "predicted"))
#      predicted
#actual  no yes
#   no  427  16
#   yes 117 195
```