Line H. Clemmensen
Section of Statistics and Data Analysis
DTU Compute

Exercises 02582
Module 9
Spring 2018

April 3, 2018

Topics: Random forest, bagging, and boosting

Resources for this exercise:

Listing 1: Resources in Matlab

```
% Random Forest:
b = TreeBagger(B,X,y,'OOBVarImp','on','Method','classification',
        'NVarToSample',20,'MinLeaf',5);
% plot out-of-bag error:
plot(1:B,oobError(b),'r')
% decide on tree model to use for bagging or boosting:
t1 = ClassificationTree.template('MinLeaf',5,'Prune','on');
% Boosting:
ens1 = fitensemble(X,y,'AdaBoostM2',B,t1,'crossval','on','LearnRate',1);
```

Listing 2: Resources in R

```
# Random Forest:
install('randomForest')
rf <- randomForest(y ~ ., data=zip, importance=TRUE,
proximity=TRUE)
# Boosting:
install('adabag')
b <- boosting(y~., data=zip, boos=TRUE, mfinal=10)
```

Listing 3: Resources in Python

```
# Random Forest:
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(max_depth=2, random_state=0)
clf.fit(X, y)
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini'
            max_depth=2, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
#Boosting:
from sklearn.ensemble import AdaBoostClassifier
bdt_real = AdaBoostClassifier(
    DecisionTreeClassifier(max_depth=2),
    n_estimators=600,
    learning_rate=1)
bdt_real.fit(X_train, y_train)
```

Line H. Clemmensen
Section of Statistics and Data Analysis
DTU Compute

Exercises 02582
Module 9
Spring 2018

April 3, 2018

Topics: Random forest, bagging, and boosting

Exercises:

1 Derive the variance for an average of B i.i.d. random variables, each with variance σ^2. What happens when B goes to infinity, i.e. we increase the number of trees?
   (a) Tip: $Var(X+Y) = Var(X) + Var(Y)$ if X and Y are independent.
   (b) Tip: $Var(aX) = a^2 Var(X)$ for a constant a.

2 Derive the variance for an average of B i.d. (not independent) random variables, each with variance σ^2, and positive pairwise correlation ρ. What happens when B goes to infinity, i.e. we increase the number of trees?
   (a) Tip: $Var(X+Y) = COV(X+Y, X+Y) = COV(X,X) + COV(X,Y) + COV(Y,X) + COV(Y,Y)$.
   (b) Tip: correlation $= \rho$ => covariance $= \rho\sigma^2$.

3 How does this relate to the following plot from ESL?

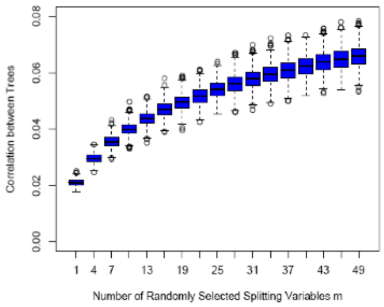

FIGURE 15.9. *Correlations between pairs of trees drawn by a random-forest regression algorithm, as a function of m. The boxplots represent the correlations at 600 randomly chosen prediction points x.*

4 Load the zip data and make a classification using random forests. As a minimum you need to tune the parameters: The number of trees, the number of variables to sample, and the tree size. Make an effort of explaining what happens when you tune each of the parameters. Which one is more important? Additionally answer:

   (a) How should you set the parameters to run bagging?

   (b) How should you set the parameters to run CART?

5 Fit bagging on the zip data. You should at least tune the number of models, as well as the individual models to obtain the best classification rates. Which tuning parameters are the most important for obtaining good performance with bagging?

6 Fit boosting on the zip data. You should tune the learning rate, the number of models, and also the individual models to obtain the best classification rates. Which tuning parameters are the most important for obtaining good performance with boosting?

End of exercise