

Module 8: Tree-based Methods

TMA4268 Statistical Learning V2020

Stefanie Muff, Department of Mathematical Sciences, NTNU

March xx, 2020

Introduction

Learning material for this module

- James et al (2013): An Introduction to Statistical Learning. Chapter 8.
- Classnotes (todo: give link, or write on board)

Some of the figures in this presentation are taken (or are inspired) from James et al. (2013).

What will you learn?

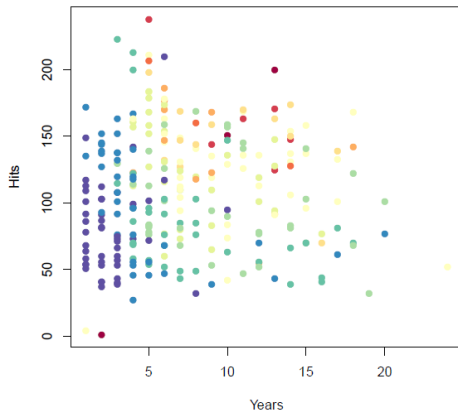
You will get to know

- Decision trees
- Regression trees
- Classification trees
- Pruning a tree
- Bagging
- Variable importance
- Random forests
- Boosting

and learn how to apply all that.

Example 1 (from chapter 8.1; Hitters data)

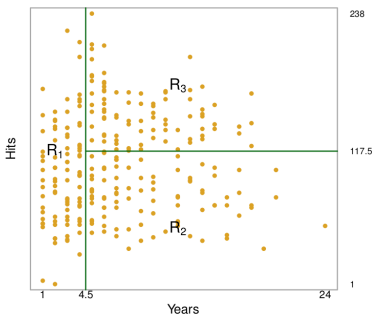
- Baseball players' salaries may depend on their experience (in years) and the number of hits.
- High salaries (yellow, red) vs low salaries (blue, green), salaries given on log-scale. How can these be stratified for prediction of the salary?



Main idea of tree-based methods

- We can divide the area into rectangles with similar salaries.
- Do this by deriving a set of decision (splitting) rules for segmenting the predictor space into a number of finer and finer regions.
- All points in the same region will be given the same predictive value (the mean of all values in that square, or a majority vote).

In two dimensions (for two variables) a three-region partition may look like this:



The series of splitting rules can be visualized with a *regression tree*.

The following tree (which corresponds to the split in the previous slide) has three *leafs* (terminal nodes), and two *internal nodes*:



More than two predictors?

- With more than two predictors we cannot draw the partition of the data in a coordinate system, but we can still draw the regression tree.
- Before we discuss how the algorithm splits the data into regions, let us look at a somewhat more interesting example.

Example 2: Detection of Minor Head Injury

(Artificial data)

- Data from patients that enter hospital. The aim is to quickly assess whether a patient has a brain injury or not.
- Patients are investigated and (possible) asked questions.
- Our job: To build a good model to predict quickly if someone has a brain injury. The method should be
 - **easy** to interpret for the medical personnel that are not skilled in statistics, and
 - **fast**, such that the medical personnel quickly can identify a patient that needs treatment.

→ This can be done by using tree-based methods.

Note: Of course, the model should be built *before* a new emergency patient arrives, using data that is already available.

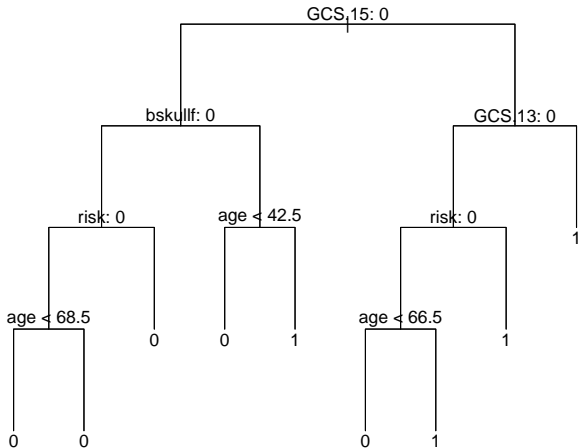
The dataset includes data about 1321 patients and is a modified and smaller version of the (simulated) dataset `headInjury` from the `DAAG` library.

```
##      amnesia bskullf GCSdecr GCS.13 GCS.15 risk consc oskullf vomit
## 3          0        0        0        0        0        0        0        0        0
## 9          0        0        0        0        0        1        0        0        0
## 11         0        0        0        0        0        0        0        0        0
## 12         1        0        0        0        0        0        0        0        0
## 14         0        0        0        0        0        0        0        0        0
## 16         0        0        0        0        0        0        0        0        0
##      brain.injury age
## 3                0 44
## 9                0 67
## 11               0 62
## 12               0  1
## 14               0 55
## 16               0 63
```

- The variable `brain.injury` will be the response of our model: It has value 1 if a person has an acute brain injury, and 0 otherwise.
- 250 (19%) of the patients have a clinically important brain injury.
- The 10 variables used as explanatory variables describe the state of the patient, for example
 - Is he/she vomiting?
 - Is the Glasgow Coma Scale (GCS) score¹ after 2 hours equal to 15 (or not)?
 - Has he/she an open skull fracture?
 - Has he/she had a loss of consciousness?
 - and so on.

¹The GCS scale goes back to an article in the Lancet in 1974, and is used to describe the level of consciousness of patients with an acute brain injury. See <https://www.glasgowcomascale.org/what-is-gcs/>

The classification tree made from a training set of 850 randomly drawn observations (training set) for the head injury example looks like this:



Note: The split criterion at each node is to the left. For example, “GCS.15:0” means that “GCS.15=0” goes left, and “GCS.15=1” goes right.

```
print(headtree)
```

```
## node), split, n, deviance, yval, (yprob)
##      * denotes terminal node
##
##  1) root 850 819.00 0 ( 0.8129 0.1871 )
##    2) GCS.15: 0 711 520.00 0 ( 0.8805 0.1195 )
##      4) bskullf: 0 663 398.00 0 ( 0.9110 0.0890 )
##        8) risk: 0 487 203.00 0 ( 0.9466 0.0534 )
##          16) age < 68.5 445 131.00 0 ( 0.9663 0.0337 ) *
##          17) age > 68.5 42  48.30 0 ( 0.7381 0.2619 ) *
##          9) risk: 1 176 170.00 0 ( 0.8125 0.1875 ) *
##        5) bskullf: 1 48  66.20 1 ( 0.4583 0.5417 )
##          10) age < 42.5 13  11.20 0 ( 0.8462 0.1538 ) *
##          11) age > 42.5 35  43.60 1 ( 0.3143 0.6857 ) *
##      3) GCS.15: 1 139 192.00 1 ( 0.4676 0.5324 )
##        6) GCS.13: 0 121 167.00 0 ( 0.5289 0.4711 )
##          12) risk: 0 78 101.00 0 ( 0.6538 0.3462 )
##            24) age < 66.5 66  77.30 0 ( 0.7273 0.2727 ) *
##            25) age > 66.5 12  13.50 1 ( 0.2500 0.7500 ) *
##          13) risk: 1 43  52.70 1 ( 0.3023 0.6977 ) *
##        7) GCS.13: 1 18   7.72 1 ( 0.0556 0.9444 ) *
```

- By using simple decision rules related to the most important explanatory variables the medical staff can now assess the probability of a brain injury.
- The decision can go “top down”, because the most informative predictors are usually split first.
- Example: The staff might check if the Glasgow Coma Scale of the patient is 15 after 2h, and if it was 13 at the beginning. In that case, the probability of brain injury is estimated to be 0.944 (node 7 in printout).

Advantages:

- Decision trees are easier to interpret than many of the classification (and regression) methods that we have studied so far.
- Decision trees provide an easy way to visualize the data for non-statisticians.

Glossary

- CARTs (Classification and regression trees) are usually drawn upside down, where the top node is called the *root*.
- The *terminal nodes* or *leaf nodes* are the nodes at the bottom, with no splitting criteria. These represent the final predicted class (for classification trees) or predicted response value (for regression trees) and are written symbolically as R_j for $j = 1, 2, \dots, J$
- The R_j will be referred to as *non-overlapping regions*.
- *Internal nodes* are all nodes between the root and the terminal nodes. These nodes correspond to the partitions of the predictor space.
- *Branches*: segment of the tree connecting the nodes.

We will consider only binary splits on one variable, but multiway splits and linear combination of variables are possible - but not so common.

Constructing a decision tree

You can construct decision trees for both classification and regression problems, first we focus on constructing a regression tree.

Regression tree (continuous outcome)

Assume that we have a dataset consisting of n pairs (\mathbf{x}_i, y_i) , $i = 1, \dots, n$, and each predictor is $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. The aim is to predict y_i .

Two steps:

1. Divide the predictor space into non-overlapping regions R_1, R_2, \dots, R_J .
2. For every observation that falls into region R_j we make the same prediction - which is the mean of the responses for the training observations that fall into R_j .

But: How to divide the predictor space into non-overlapping regions R_1, R_2, \dots, R_J ?

We could try to minimize the RSS (residual sums of squares) on the training set given by

$$\text{RSS} = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

where \hat{y}_{R_j} is the mean response for the training observations in region j . The mean \hat{y}_{R_j} is also the predicted value for a new observations that falls into region j .

To do this we need to consider every partition of the predictor space, and compute the RSS for each partition.

But: An exhaustive search over possible splits is *computationally infeasible*! In fact, constructing optimal binary decision trees is an NP-complete problem (Hyafil and Rivest 1976).

Recursive binary splitting

- A *greedy* approach is taken (aka top-down) - called *recursive binary splitting*. The idea is to find a split that minimizes RSS at each step.
- We start at the top of the tree and divide the predictor space into two regions, R_1 and R_2 by making a decision rule for one of the predictors x_1, x_2, \dots, x_p . If we define the two regions by $R_1(j, s) = \{x \mid x_j < s\}$ and $R_2(j, s) = \{x \mid x_j \geq s\}$, it means that we need to find the (predictor) j and (splitting point) s that minimize

$$\sum_{i: x_i \in R_1(j, s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j, s)} (y_i - \hat{y}_{R_2})^2 ,$$

where \hat{y}_{R_1} and \hat{y}_{R_2} are the mean responses for the training observations in $R_1(j, s)$ and $R_2(j, s)$ respectively. This way we get the two first branches in our decision tree.

- We repeat the process to make branches further down in the tree.
- For every iteration we let each single split depend on *only one of the predictors*, giving us two new branches.
- This is done *successively* and in each step we choose the split that gives the *best split at that particular step*, i.e., the split that gives the smallest RSS.
- Continue splitting the predictor space until we reach some *stopping criterion*. For example we stop when a region contains less than 10 observations or when the reduction in the RSS is smaller than a specified limit.

Q: Why is this algorithm called *greedy*?

Regression tree: ozone example

Consider the `ozone` data set from the `ElemStatLearn` library. The data set consists of 111 observations on the following variables:

- `ozone` : the concentration of ozone in ppb
- `radiation`: the solar radiation (langleys)
- `temperature` : the daily maximum temperature in degrees F
- `wind` : wind speed in mph

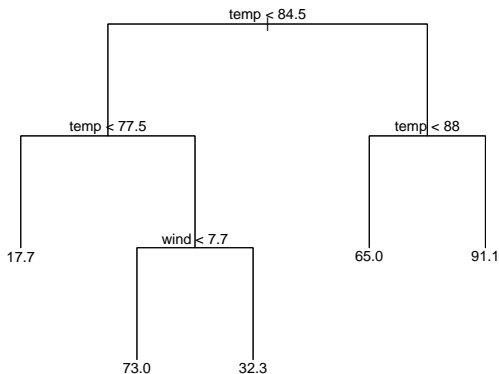
```
head(myozone)
```

```
##      ozone radi temp wind
## 1      41  190   67  7.4
## 2      36  118   72  8.0
## 3      12  149   74 12.6
## 4      18  313   62 11.5
## 5      23  299   65  8.6
## 6      19   99   59 13.8
```

Suppose we want to get an estimate of the ozone concentration based on the measurement of wind speed and the daily maximum temperature.

→ We can fit a regression tree to the data, with **ozone** as our response variable and **temperature** and **wind** as predictors (not including radiation to make this easier to see). This gives us the following regression tree.

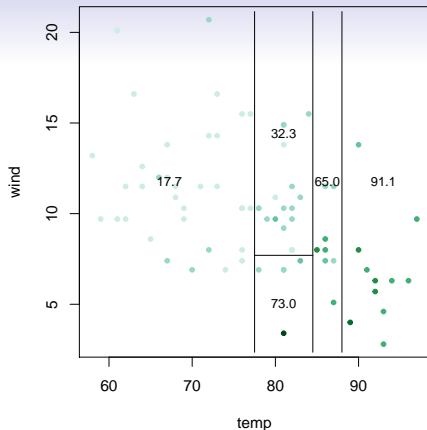
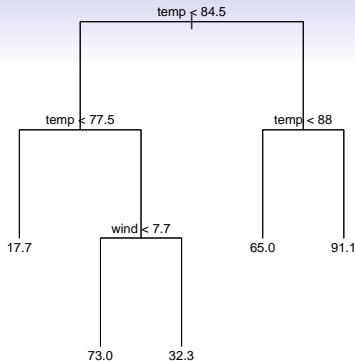
```
ozone.trainID = sample(1:111, 75)
ozone.train = myozone[ozone.trainID, ]
ozone.test = myozone[-ozone.trainID, ]
# use the default settings in the tree function:
ozone.tree = tree(ozone ~ temp + wind, data = ozone.train)
```



- We see that **temperature** is the “most important” predictor for predicting the ozone concentration. Observe that we can split on the same variable several times.
- Focus on the regions R_j , $j = 1, \dots, J$. What is J here? Answer: $J = 6$ (= the number of leaf nodes).

Below we see the partition of the **ozone** data set given in the `ozone.tree`, where

- the ozone levels have been color-coded, where a darker color corresponds to a higher ozone concentration.
- each rectangle corresponds to one leaf node.
- each number corresponding to a leaf node has been found by taking an average of all observations (in the training set) in the corresponding region (rectangle).
- each partition line corresponds to one internal node of a binary partition.



Q:

- Explain the connection between the tree and the region plot.
- Why is recursive binary splitting classified as a greedy algorithm?
- Discuss the advantages and disadvantages of letting each single split depend on only one of the predictors.
- Does our tree automatically include interactions between variables?

A:

- Each leaf node corresponds to one region with prediction \hat{y}_{R_j} , and each internal node corresponds to a vertical or a horizontal line.
- Recursive splitting does not necessarily give the optimal global solution, but will give the best solution at each split (given what is done previously).
- If the true connection between the response ozone and temperature and wind was so that splits should have been made on the diagonal of the temperature and wind space, that would take many splits on each of temperature and wind to produce. (See more below where we ask the same question for the classification setting.)
- Yes, we may have different effect of wind for different values of temperature.

R: function `tree` in library `tree`

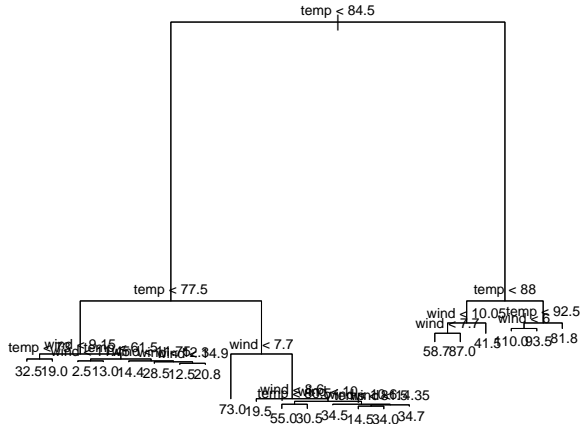
- R package `tree` by Brian D. Ripley (B. Ripley 2019).
- You get information about the function when you type `?tree` into the R console.
- Note: The default choice for a function to minimize is the deviance, and for normal data (as we may assume for regression), the deviance is proportional to the RSS.
- A competing R function is `rpart`, explained in <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf>

Stopping criterion

- When building the tree, we can (in principle) split until each leaf corresponds to one data point.
- Usually we use a less stringent stopping criterion, like the minimal number of nodes per region and/or the minimum reduction in the RSS.
- For example, the default in `tree()`: is given by `mincut=5`, `minsize=10`, `mindev=0.01`, thus a minimal size of 5 nodes and the minimal reduction in deviance (RSS) equal to 0.01.
- We could make the ozone tree much deeper by decreasing the minimal size of a region and changing the minimum reduction in RSS:

```
ozone.tree2 = tree(ozone ~ temp + wind, data = ozone.train, control = tree.control(75,  
  mincut = 2, minsize = 4, mindev = 0.001))
```

With the above command, the tree would become very fine. Overfitting?



Tree performance

To test the predictive performance of our regression tree, we have randomly divided our observations into a test and a training set (here 1/3 test). For the two trees (the shallow and the deep one from above), we get the following test MSEs:

```
ozone.pred = predict(ozone.tree, newdata = ozone.test)
ozone.MSE = mean((ozone.pred - ozone.test$ozone)^2)
ozone.MSE
```

```
## [1] 384.491
```

```
ozone.pred2 = predict(ozone.tree2, newdata = ozone.test)
ozone.MSE2 = mean((ozone.pred2 - ozone.test$ozone)^2)
ozone.MSE2
```

```
## [1] 491.266
```

How do we know if our tree performs good, or if there would be trees with a better predictive performance?

Pruning

- If we have a data set with many predictors or choose a stringent stopping criterion, we may fit a (too) large tree. Thus, the number of observations from the training set that falls into some of the regions R_j may be small \rightarrow *overfitting*?
- A smaller tree with fewer splits leads to fewer regions R_1, \dots, R_J with more observations. This might lead to lower variance and better interpretation at the cost of more bias.
- One possible alternative to the process described above is to grow the tree only so long as the decrease in the RSS due to each split exceeds some (high) threshold.
- This strategy will result in smaller trees, but is too short-sighted: a seemingly worthless split early on in the tree might be followed by a very good split — that is, a split that leads to a large reduction in RSS later on.

Cost complexity pruning

- Better idea: to grow a very large tree T_0 , and then *prune* it back in order to obtain a subtree
- *Cost complexity pruning* is used for this: We try to find a subtree $T \subset T_0$ that (for a given value of α) minimizes

$$C_\alpha(T) = Q(T) + \alpha|T|,$$

where $Q(T)$ is our cost function, $|T|$ is the number of terminal nodes in tree T . The parameter α is then a parameter penalizing the number of terminal nodes, ensuring that the tree does not get too many branches.

- For regression trees we choose

$$Q(T) = \sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2,$$

and adapted cost functions for classification trees (the entropy (deviance), Gini or misclassification rate, see below).

- Given a value of α we get a pruned tree (note: the same pruned tree within a small range of α).
- For $\alpha = 0$ we get T_0 and as α increases we get shorter and shorter trees.

So, which value of α is best?

- We can use K -fold cross-validation to find out!
- Importantly, by increasing α , branches get pruned in a *hierarchical (nested) fashion*.

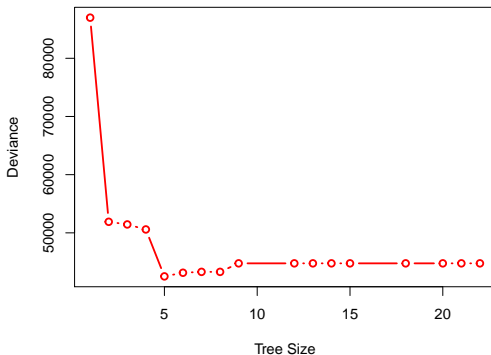
Please study this [note from Bo Lindqvist in MA8701 in 2017 - Advanced topics in Statistical Learning and Inference](#) for an example of how we perform cost complexity pruning in detail. Alternatively, this method, with proofs, are given in B. D. Ripley (1996), Section 7.2.

Building a regression tree: Algorithm 8.1

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α .
3. Use K -fold cross-validation to choose α . That is, divide the training observations into k folds. For each $k = 1, \dots, K$:
 - Repeat Steps 1 and 2 on all but the k th fold of the training data.
 - Evaluate the mean squared prediction error on the data in the left-out k th fold, as a function of α .
 - Average the results for each value of α , and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α .

Pruning the ozone tree

Let us start with the tree with many leaves, and then prune it with 5-fold CV. We can then plot the CV error as a function of tree size (instead of α – why?):



Interestingly, it seems like a tree with 5 leaves performs best - which corresponds to the original choice.

Classification trees (binary or categorical outcome)

- Remember the minor head injury example: Binary outcome (disease yes/no).
- Now allow for $K \geq 2$ number of classes for the response.
- Building a decision tree in this setting is similar to building a regression tree for a quantitative response, but there are two main differences: *the prediction* and *the splitting criterion*.

1) The prediction:

- In the regression case we use the mean value of the responses in R_j as a prediction for an observation that falls into region R_j .
- For the *classification case*, however, we have two possibilities:
 - Majority vote: Predict that the observation belongs to the most commonly occurring class of the training observations in R_j .
 - Estimate the probability that an observation x_i belongs to a class k , $\hat{p}_{jk}(x_i)$, and then classify according to a threshold value. This estimated probability is the proportion of class k training observations in region R_j , with n_{jk} observations. Region j has N_j observations.

$$\hat{p}_{jk} = \frac{1}{N_j} \sum_{i: x_i \in R_j} I(y_i = k) = \frac{n_{jk}}{N_j}.$$

2) The splitting criterion: We do not use RSS as a splitting criterion for a qualitative variable. Instead we can use some *measure of impurity* of the node. For leaf node j and class $k = 1, \dots, K$:

- **Gini index:**

$$G = \sum_{k=1}^K \hat{p}_{jk}(1 - \hat{p}_{jk}) ,$$

which is small if all of the \hat{p}_{jk} 's are close to 0 or 1.

- **Cross entropy:**

$$D = - \sum_{k=1}^K \hat{p}_{jk} \log \hat{p}_{jk} .$$

Here \hat{p}_{jk} is the proportion of training observation in region j that are from class k .

When making a split in our classification tree, we want to minimize the Gini index or the cross-entropy.

Why don't we just minimize the misclassification error

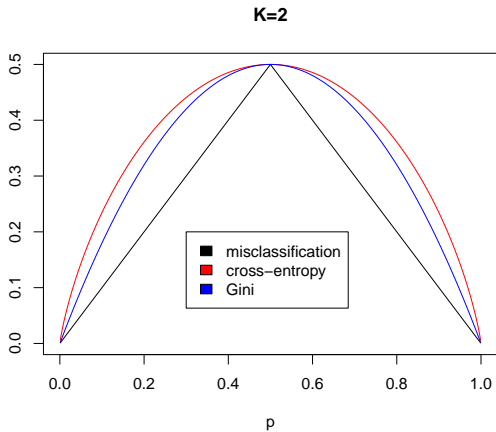
$$E = 1 - \max_k \hat{p}_{jk} ?$$

A:

- The Gini index and Entropy are *measure of impurity*. They are more sensitive to changes in the node probabilities.
- Example for two classes: Assume we have two classes with 400 nodes each, written as (400, 400). Now we can choose between two splits:
 - Split 1: (100,300) and (300,100)
 - Split 2: (200,400) and (200,0).

Both splits result in 25% misclassification. Which of these splits is better? Probably the second one, because it produces a *pure node*.

- Moreover, The Gini index and Entropy are differentiable (preferred for numerical optimization!)



Minor head injury - split with cross-entropy

We use the 850 training samples to get a tree using cross-entropy (deviance):

```
tree.HIClass = tree(brain.injury ~ ., data = headInjury2, subset = train,  
  split = "deviance")  
summary(tree.HIClass)
```

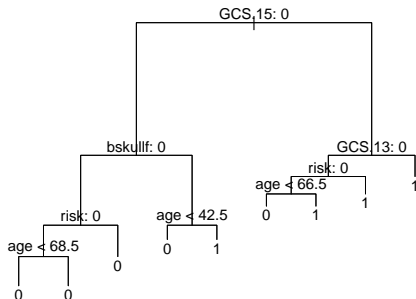
```
##  
## Classification tree:  
## tree(formula = brain.injury ~ ., data = headInjury2, subset = train,  
##      split = "deviance")  
## Variables actually used in tree construction:  
## [1] "GCS.15" "bskullf" "risk" "age" "GCS.13"  
## Number of terminal nodes: 9  
## Residual mean deviance: 0.66 = 555 / 841  
## Misclassification error rate: 0.126 = 107 / 850
```

Remark: the deviance is a scaled version of the cross entropy:

$$-2 \sum_{k=1}^K n_{jk} \log \hat{p}_{jk} \text{ , where } \hat{p}_{jk} = \frac{n_{jk}}{N_j} \text{ ,}$$

thus `split="deviance"` implies that we split according to the entropy criterion.

```
plot(tree.HIClass, type = "proportional")  
text(tree.HIClass, pretty = 1)
```



- With `type="proportional"`, the length of branches are now proportional to the decrease in impurity.
- The classification tree has two terminal nodes with factor “0” originating from the same branch. Why do we get this “unnecessary” split? (Hint: Node purity)

Minor head injury - split with Gini index

The same analysis with the Gini index:

```
tree.HIClassG = tree(brain.injury ~ ., headInjury2, subset = train,  
  split = "gini")  
summary(tree.HIClassG)
```

```
##  
## Classification tree:  
## tree(formula = brain.injury ~ ., data = headInjury2, subset = train,  
##   split = "gini")  
## Variables actually used in tree construction:  
## [1] "GCS.15" "bskullf" "risk" "age" "oskullf" "vomit" "amnesia"  
## [8] "consc" "GCS.13"  
## Number of terminal nodes: 75  
## Residual mean deviance: 0.52 = 403 / 775  
## Misclassification error rate: 0.109 = 93 / 850
```


Checking predictions

We also use the classification tree to predict the status of the patients in the test set.

With the deviance:

```
library(caret)
tree.pred = predict(tree.HIClass, headInjury2[test, ], type = "class")
(confMat <- confusionMatrix(tree.pred, reference = headInjury2[test,
]$brain.injury)$table)
```

```
##           Reference
## Prediction    0    1
##           0 356  42
##           1  24  49
```

```
1 - sum(diag(confMat))/sum(confMat[1:2, 1:2])
```

```
## [1] 0.140127
```

With the Gini-index:

```
tree.predG = predict(tree.HIClassG, headInjury2[test, ], type = "class")
(confMatG <- confusionMatrix(tree.predG, reference = headInjury2[test,
]$brain.injury)$table)
```

```
##           Reference
## Prediction    0    1
##           0 354  47
##           1  26  44
```

```
1 - sum(diag(confMatG))/sum(confMatG[1:2, 1:2])
```

```
## [1] 0.154989
```

Study the confusion matrices on the previous slide:

- Which algorithm makes more errors?
- Is one type of mistakes more severe than the other?
- Discuss if/how it is possible to change the algorithm in order to decrease the number of severe mistakes.

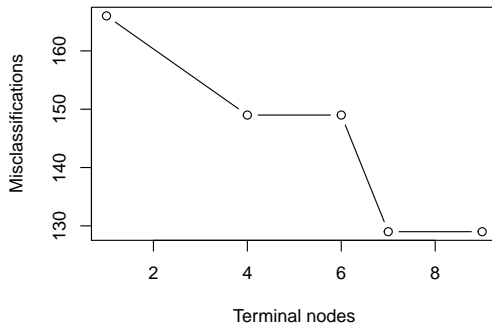
Building a classification tree: Algorithm 8.1

- Like for regression trees, we would like to find classification trees that are good at prediction.
- Algorithm 8.1 can be used with misclassification, the Gini index, or cross-entropy instead of the RSS as quality measure.

Finding an optimal classification tree

We prune the classification tree that was built with the deviance criterion. We use the misclassification error to do so:

```
set.seed(1)
cv.head = cv.tree(tree.HIClass, FUN = prune.misclass)
plot(cv.head$size, cv.head$dev, type = "b", xlab = "Terminal nodes",
     ylab = "Misclassifications")
```



(Todo: Exercise with `cv.head=cv.tree(tree.HIClass,FUN=prune.tree)` , that is, pruning according to deviance;
Note: I think it is not possible to prune according to the Gini index - why??)

The function `cv.tree` automatically does 10-fold cross-validation. `dev` is here the number of misclassifications.

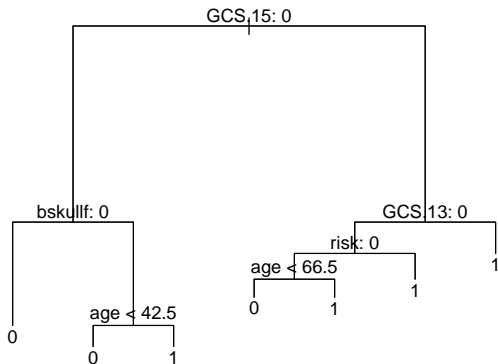
```
print(cv.head)
```

```
## $size
## [1] 9 7 6 4 1
##
## $dev
## [1] 129 129 149 149 166
##
## $k
## [1] -Inf 0.0 6.0 6.5 11.0
##
## $method
## [1] "misclass"
##
## attr("class")
## [1] "prune"          "tree.sequence"
```


- We have done cross-validation on our training set of 850 observations. According to the plot, the number of misclassifications is as low for 7, 8 or 9 nodes, so we choose 7 terminal nodes as the smallest model (lowest variance).
- We prune the classification tree according to this value:

```
prune.HIClass = prune.misclass(tree.HIClass, best = 7)
```

```
plot(prune.HIClass)
text(prune.HIClass, pretty = 1)
```



→ No unnecessary splits left, and we have a simple and interpretable decision tree.

How is the predictive performance of the model affected?

```
tree.pred.prune <- predict(prune.HIClass, headInjury2[test, ], type = "class")
(confMat <- confusionMatrix(tree.pred.prune, headInjury2[test, ]$brain.injury)$table)
```

```
##           Reference
## Prediction    0    1
##           0 356  42
##           1  24  49
```

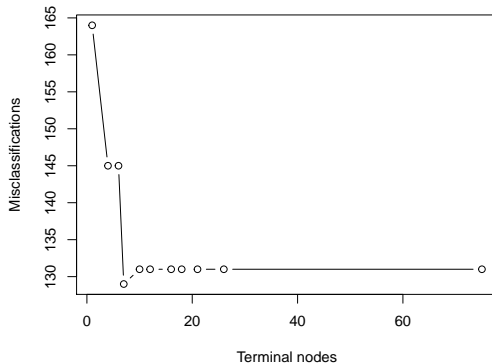
```
1 - (sum(diag(confMat))/sum(confMat[1:2, 1:2]))
```

```
## [1] 0.140127
```

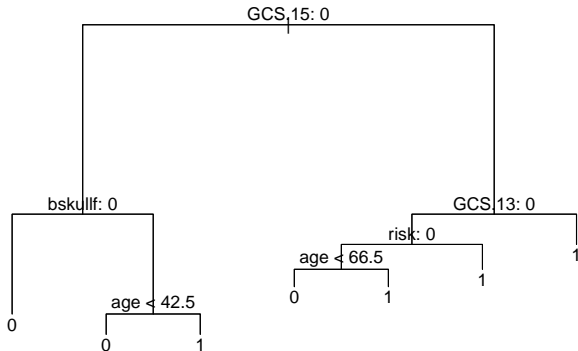
We see that the misclassification rate is as small as before, thus the pruned tree is as good as the original tree for the test data.

The same repeated for the Gini-grown tree - comment on what is done.

```
set.seed(10)
cv.headG = cv.tree(tree.HIClassG, FUN = prune.misclass)
plot(cv.headG$size, cv.headG$dev, type = "b", xlab = "Terminal nodes",
     ylab = "Misclassifications")
```



```
prune.HIClassG = prune.misclass(tree.HIClassG, best = 7)
plot(prune.HIClassG)
text(prune.HIClassG, pretty = 1)
```



Compare this tree to the one from pruning the deviance-based tree.

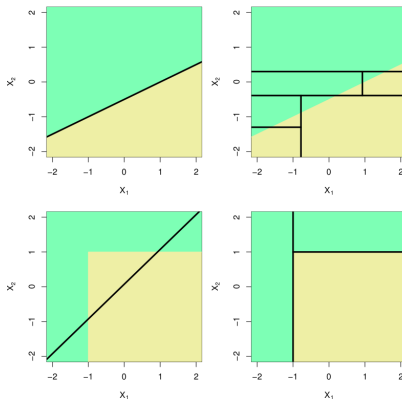
Questions:

Discuss the bias-variance tradeoff of a regression tree when increasing/decreasing the number of terminal nodes, i.e:

- What happens to the bias?
- What happens to the variance of a prediction if we reduce the tree size?

Trees versus linear models

- What if we have x_1 and x_2 and the true class boundary (two classes) is linear in x_1, x_2 space. How can we do that with our binary recursive splits?
- What about a rectangular boundary?



From trees to forests

Advantages (+)

- Trees automatically select variables.
- Tree-growing algorithms scale well to large n , growing a tree greedily.
- Trees can handle mixed features (continuous, categorical) seamlessly, and can deal with missing data.
- Small trees are easy to interpret and explain to people.
- Some believe that decision trees mirror human decision making.
- Trees can be displayed graphically.

Disadvantages (-)

- Large trees are not easy to interpret.
- Trees do not generally have good prediction performance (high variance).
- Trees are not very robust, a small change in the data may cause a large change in the final estimated tree.

What is next?

Several of the above listed disadvantages can be addressed by

- **Bagging:** grow many trees (from bootstrapped data) and average - to get rid of the non-robustness and high variance by averaging.
- **Random forests:** inject more randomness (and even less variance) by just allowing a random selection of predictors to be used for the splits at each node.
- **Boosting:** make one tree, then another based on the residuals from the previous, repeat. The final predictor is a weighted sum of these trees.

In addition:

- **Variable importance** - to see which variables make a difference (now that we have many trees).

But first:

Leo Breiman - the inventor of CART, bagging and random forests

Quotation from [Wikipedia](#)

Leo Breiman (January 27, 1928 – July 5, 2005) was a distinguished statistician at the University of California, Berkeley. He was the recipient of numerous honors and awards, and was a member of the United States National Academy of Science. Breiman's work helped to bridge the gap between statistics and computer science, particularly in the field of machine learning. His most important contributions were his work on classification and regression trees and ensembles of trees fit to bootstrap samples. Bootstrap aggregation was given the name bagging by Breiman. Another of Breiman's ensemble approaches is the random forest.

From [Breimans obituary](#)

BERKELEY – Leo Breiman, professor emeritus of statistics at the University of California, Berkeley.

"It is trite to say so, but Leo Breiman was indeed a Renaissance man, and we shall miss him greatly," said Peter Bickel, professor of statistics and chairman this summer of UC Berkeley's statistics department.

Breiman retired in 1993, but as a Professor in the Graduate School, he continued to get substantial National Science Foundation grants and supervised three Ph.D. students. Bickel said that some of Breiman's best work was done after retirement.

"In particular," said Bickel, "he developed one of the most successful state-of-the-art classification programs, 'Random Forest.' This method was based on a series of new ideas that he developed in papers during the last seven years, and it is extensively used in government and industry."

Breiman's best known work is considered to be "Classification and Regression Trees," a work in collaboration with three other scholars that facilitates practical applications, such as the diagnosis of diseases, from a multitude of symptoms.

Bagging

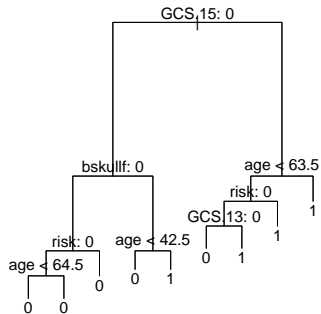
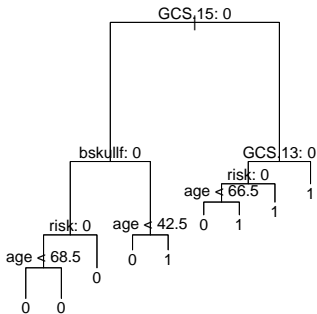
- Decision trees often suffer from high variance. By this we mean that the trees are sensitive to small changes in the predictors: If we change the observation set, we may get a very different tree.
- Another way to understand “high variance” is that, if we split our training data into two parts and fit a tree on each, we might get rather different decision trees.

Let's draw a new training set for our data and see what happens if we fit our full classification tree (deviance grown).

```

set.seed(33)
N = dim(headInjury2)[1]
train2 = sample(1:N, 850) # Draw a new training sample (same size).
tree.HIClass2 = tree(brain.injury ~ ., data = headInjury2, subset = train2,
  split = "deviance")
par(mfrow = c(1, 2))
plot(tree.HIClass)
text(tree.HIClass, pretty = 0)
plot(tree.HIClass2)
text(tree.HIClass2, pretty = 0)

```



This classification tree is constructed by using 850 observations, just like the tree in the classification trees section, but we get two different trees that will give different predictions for a test set.

To reduce the variance of decision trees we can apply *bootstrap aggregating (bagging)*, invented by Leo Breiman in 1996 (Breiman 1996).

Recall: Variance for independent datasets

- Assume we have B i.i.d. observations of a random variable X each with the same mean and with variance σ^2 . We calculate the mean $\bar{X} = \frac{1}{B} \sum_{b=1}^B X_b$. The variance of the mean is

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{B} \sum_{b=1}^B X_b\right) = \frac{1}{B^2} \sum_{b=1}^B \text{Var}(X_b) = \frac{\sigma^2}{B}.$$

By averaging we get reduced variance. This is the basic idea!

- For decision trees, if we have B training sets, we could estimate $\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \dots, \hat{f}_B(\mathbf{x})$ and average them as

$$\hat{f}_{avg}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{f}_b(\mathbf{x}) .$$

However, we do not have many independent data set - so we use *bootstrapping* to construct B data sets.

Recall: Bootstrapping

- Problem: we want to draw samples from a population with distribution f .
- But: we do not know f and do not have a population to draw from, we only have our one sample.
- Solution: we may use our sample as an empirical estimate for the distribution f - by assuming that each sample point has probability $1/n$ for being drawn.
- Therefore: we draw *with replacement* n observations from our sample - and that is our first *bootstrap sample*.
- We repeat this B times and get B bootstrap samples - that we use as our B data sets.

Bagging = Bootstrap aggregating

- Idea: For each bootstrap sample $b = 1, \dots, B$ we construct a decision tree, $\hat{f}^{*b}(x)$, and we then use information from all of the trees to draw inference.
- For a *regression tree*, we take the average of all of the predictions and use this as the final result:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x).$$

- For a *classification tree* we record the predicted class (for a given observation x) for each of the B trees and use the most occurring classification (*majority vote*) as the final prediction - or alternatively average posterior probabilities for each class.

- Originally, Breiman (1996) suggested to prune each tree, but later research has found that it is better to leave the trees at maximal size (a bushy tree), to make the trees as different from each other as possible.
- The number B is chosen to be as large as “necessary”. An increase in B will not lead to overfitting, and B is not regarded as a tuning parameter. If a goodness of fit measure is plotted as a function of B (soon) we see that (given that B is large enough) increasing B will not change the goodness of fit measure.
- To find out which number B is sufficient, we do *not* need to run cross-validation, but can do something smarter (next slide).

Out-of-bag error estimation

- We use a subset of the observations in each bootstrap sample. Recall (module 5) that the probability that an observation is in the bootstrap sample is approximately $1 - e^{-1} = 0.6321206$ ($\approx 2/3$).
- When an observation is left out of the bootstrap sample it is not used to build the tree, and we can use this observation as a part of a “test set” to measure the predictive performance and error of the fitted model, $\hat{f}^{*b}(x)$.

Terminology:

- The observations left out are referred to as the *out-of-bag* (OOB) observations.
- The measured error of the $B/3$ predictions is called the *out-of-bag error*.

Example

We can do bagging by using the function `randomForest()` in the `randomForest` library.

```
library(randomForest)
set.seed(1)
r.brain.bag <- randomForest(brain.injury ~ ., data = headInjury2,
  subset = train, mtry = 10, ntree = 500, importance = TRUE)
r.brain.bag$confusion

##      0  1 class.error
## 0 637 54  0.0781476
## 1  76 83  0.4779874

1 - sum(diag(r.brain.bag$confusion))/sum(r.brain.bag$confusion[1:2,
  1:2])

## [1] 0.152941
```

The variable `mtry=10` because we want to consider all 10 predictors in each split of the tree. The variable `ntree=500` because we want to average over 500 trees.

Todo (Exercise): Make a similar example, and let students plot the error rate against `ntree` (e.g. from 1 to 500, like figure 8.8).

Predictive performance of the bagged tree on unseen test data:

```
yhat.bag = predict(r.brain.bag, newdata = headInjury2[test, ])  
misclass.bag = table(yhat.bag, headInjury2[test, ]$brain.injury)  
print(misclass.bag)
```

```
##  
## yhat.bag    0    1  
##           0 346  39  
##           1  34  52
```

```
1 - sum(diag(misclass.bag))/(sum(misclass.bag))
```

```
## [1] 0.154989
```

Note: The misclassification rate has increased slightly for the bagged tree (as compared to our previous full and pruned tree). **Typically, we would expect an improvement!**

Prediction by majority vote vs. by averaging the probabilities

- Consider the case when you have grown B classification tree with a binary response with classes 0 and 1. You might wonder which approach to choose to make a final prediction: majority vote or an average of the probabilities? Or would the prediction be the same in each case?
- The difference between these two procedures can be compared to the difference between the mean value and median of a set of numbers. If we average the probabilities and make a classification thereafter, we have the mean value. If we sort all of our classifications, so that the classifications corresponding to one class would be lined up after each other, followed by the classifications corresponding to the other class we obtain the median value.

We examine this by an example:

Suppose we have $B = 5$ (no, B should be higher - this is only for illustration) classification tree and have obtained the following 5 estimated probabilities: $\{0.4, 0.4, 0.4, 0.4, 0.9\}$. If we average the probabilities, we get 0.5, and if we use a cut-off value of 0.5, our predicted class is 1. However, if we take a majority vote, using the same cut of value, the predicted classes will be $\{0, 0, 0, 0, 1\}$. The predicted class, based on a majority vote, would accordingly be 0.

The two procedures thus have their pros and cons: By averaging the predictions no information is lost. We do not only get the final classification, but the probability for belonging to the class 0 or 1. However, this method is not robust to outliers. By taking a majority vote, outliers have a smaller influence on the result.

When should we use bagging?

Bagging can also be used for prediction methods (regression and classification) that are not trees.

- According to Breiman (1996) Bagging can improve accuracy when,
 - the vital element is the instability of the prediction method.
 - perturbing the learning set can cause significant changes in the predictor constructed.

Breiman (1996) suggests that these methods should be suitable for bagging:

- Neural nets, classification and regression trees, subset selection in linear regression.

However, bagging is not suitable for nearest neighbour classification, since the stability of these methods with respect to perturbations of the data distinguishes them from competitors such as trees and neural nets.

Variable importance plots

- Bagging typically (but not always) improves accuracy of predictions.
- Drawback: It becomes difficult to interpret the results. Instead of having one tree, the resulting model consists of many trees. So, which predictors are (how) important?
- *Variable importance plots* show *the relative importance of the predictors*: the predictors are sorted according to their importance, such that the top variables have a higher importance than the bottom variables.
- There are in general two types of variable importance plots:
 - variable importance based on decrease in node impurity.
 - variable importance based on randomization.

Variable importance based on node impurity

The term *important* relates to *total decrease in the node impurity, over splits for a predictor*, and is defined differently for regression trees and classification trees.

Regression trees:

- The importance of each predictor is calculated using the RSS.
- The algorithm records the total amount that the RSS is decreased due to splits for each predictor (there may be many splits for one predictor for each tree).
- This decrease in RSS is then averaged over the B trees. The higher the decrease, the more important the predictor.

Classification trees:

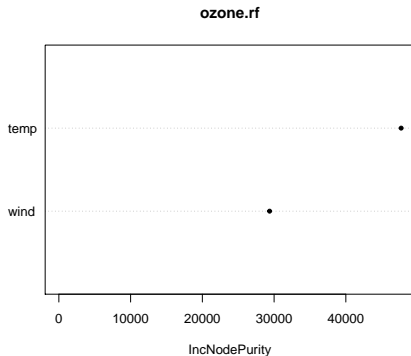
- The importance of each predictor is calculated using the Gini index.
- The importance is the mean decrease (over all B trees) in the Gini index by splits of a predictor.

Doing this in R

`varImpPlot` (or `importance`) in `randomForest` with `type=2`.

Example 1: Ozone data (regression tree)

```
set.seed(4268)
ozone.rf <- randomForest(ozone ~ temp + wind, data = ozone.train,
  ntree = 500, mtry = 2, keep.forest = FALSE, importance = TRUE)
varImpPlot(ozone.rf, type = 2, pch = 20)
```

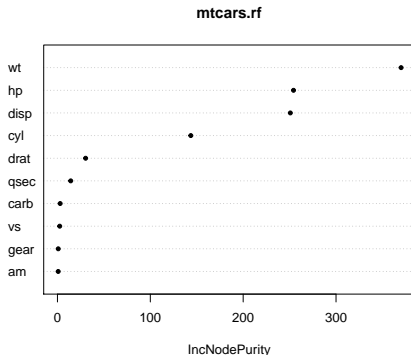


This is a bit boring (only two predictors). More interesting example?

Example 2: Auto data (regression tree)

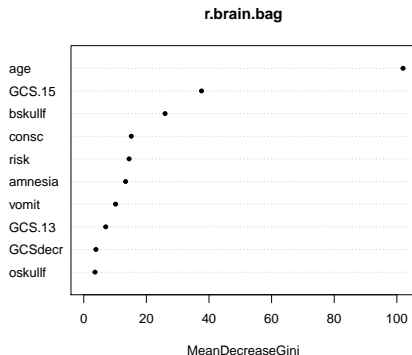
The data relates fuel consumption (mpg) to 10 aspects of automobile desing and performance.

```
data(mtcars)
mtcars.rf <- randomForest(mpg ~ ., data = mtcars, ntree = 1000, keep.forest = FALSE,
  mtry = 10, importance = TRUE)
varImpPlot(mtcars.rf, type = 2, pch = 20)
```



Example 3: Head injury (classification tree)

```
varImpPlot(r.brain.bag, type = 2, pch = 20)
```



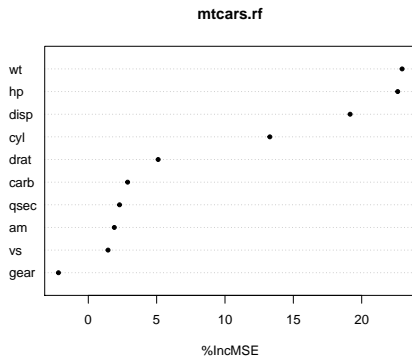
Another approach: Variable importance based on randomization

Variable importance based on randomization is calculated using the OOB sample.

- Computations are carried out for one bootstrap sample at a time.
- Each time a tree is grown the OOB sample is used to test the predictive power of the tree.
- Then for one predictor at a time, repeat the following:
 - permute the OOB observations for the j th variable x_j and calculate the new OOB error.
 - If x_j is important, permuting its observations will decrease the predictive performance.
- The difference between the two is averaged over all trees (and normalized by the standard deviation of the differences).

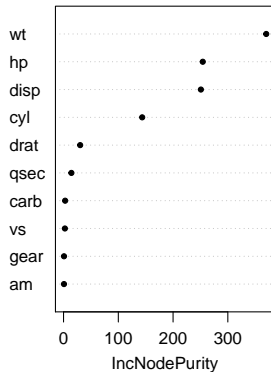
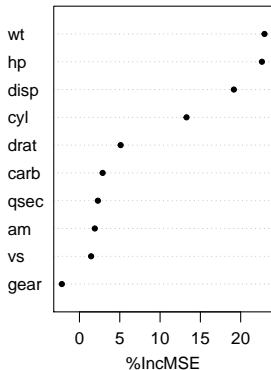
R: `varImpPlot` (or `importance`) in `randomForest` with `type=1`.


```
varImpPlot(mtcars.rf, type = 1, pch = 20)
```



The two types together: do they agree?

```
varImpPlot(mtcars.rf, pch = 20, main = "")
```



Random Forests

- If there is a strong predictor in the dataset, the decision trees produced by each of the bootstrap samples in the bagging algorithm becomes very similar: Most of the trees will use the same strong predictor in the top split.
- We have seen this for our example trees for the minor head injury example, the predictor `GCS.15` was chosen in the top split every time. This is probably the case for a large amount of the bagged trees as well.
- Not optimal, as we get B trees that are highly correlated. We don't get a large reduction in variance by averaging $\hat{f}^{*b}(x)$ when the correlation between the trees is high. In the previous section we actually even saw a (marginal) decrease in the predictive performance for the bagged tree compared to the pruned tree and the full tree.

The effect of correlation on the variance of the mean

- The variance of the average of B observations of *i.i.d.* random variables X , each with variance σ^2 is $\frac{\sigma^2}{B}$. Now, suppose we have B observations of a random variable X which are identically distributed, each with mean μ and variance σ^2 , but not independent.
- That is, suppose the variables have a positive correlation ρ

$$\text{Cov}(X_i, X_j) = \rho\sigma^2, \quad i \neq j.$$

The variance of the average is

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{B} \sum_{i=1}^B X_i\right) \\&= \sum_{i=1}^B \frac{1}{B^2} \text{Var}(X_i) + 2 \sum_{i=2}^B \sum_{j=1}^{i-1} \frac{1}{B} \frac{1}{B} \text{Cov}(X_i, X_j) \\&= \frac{1}{B} \sigma^2 + 2 \frac{B(B-1)}{2} \frac{1}{B^2} \rho \sigma^2 \\&= \frac{1}{B} \sigma^2 + \rho \sigma^2 - \frac{1}{B} \rho \sigma^2 \\&= \rho \sigma^2 + \frac{1-\rho}{B} \sigma^2 \\&= \frac{1 - (1-B)\rho}{B} \sigma^2\end{aligned}$$

Check: $\rho = 0$ and $\rho = 1$? (Most negative values of ρ will not give a positive definite covariance matrix. The covariance matrix is positive definite if $\rho > -1/(B-1)$.)

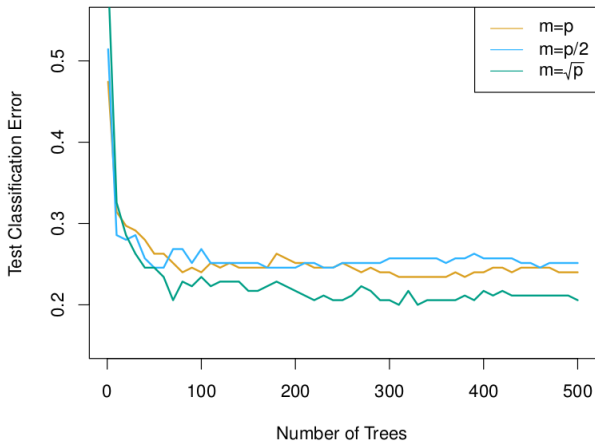
- *Random forests* provide an improvement over bagged trees by a small tweak that decorrelates the trees. This reduces the variance when we average the trees.
- As in bagging, we build a number of decision trees on bootstrapped training samples.
- But when building these decision trees, each time a split in a tree is considered, a *random selection* of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.
- A fresh selection of m predictors is taken at each split, typically
 - $m \approx \sqrt{p}$ (classification),
 - $m = p/3$ (regression).

The general idea is that for very correlated predictors m is chosen to be small.

The number of trees, B , is not a tuning parameter, and the best is to choose it large enough.

If B is sufficiently large (three times the number needed for the random forest to stabilize), the OOB error estimate is equivalent to LOOCV (see Efron and Hastie 2016, 330).

ISLR Figure 8.10, gene expression data set with 15 classes and 500 predictors:



Example

We decorrelate the brain injury classification trees by using the `randomForest()` function again, but this time we set `mtry=3` (instead for `mtry=10`). This means that the algorithm only considers three of the predictors in each split. We choose 3 because we have 10 predictors in total and $\sqrt{10} \approx 3$.

```
set.seed(1)

r.brain.rf = randomForest(brain.injury ~ ., data = headInjury2, subset = train,
  mtry = 3, ntree = 500, importance = TRUE)
```

We check the predictive performance as before:

```
r.brain.rf$confusion
```

```
##      0  1 class.error
## 0 662 29   0.0419682
## 1  93 66   0.5849057
```

```
1 - sum(diag(r.brain.rf$confusion[1:2, 1:2]))/(sum(r.brain.rf$confusion[1:2,
1:2]))
```

```
## [1] 0.143529
```

```
yhat.rf = predict(r.brain.rf, newdata = headInjury2[test, ])
```

```
misclass.rf = table(yhat.rf, headInjury2[test, ]$brain.injury)
print(misclass.rf)
```

```
##
## yhat.rf   0   1
##          0 364 47
##          1  16 44
```

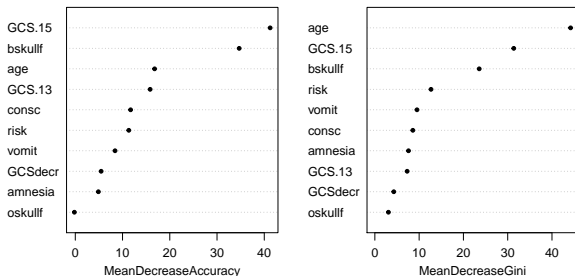
```
1 - sum(diag(misclass.rf))/(sum(misclass.rf))
```

```
## [1] 0.133758
```

The misclassification rate is slightly decreased compared to the bagged tree (and to the pruned tree).

By using the `varImpPlot()` function we can study the importance of each predictor.

```
varImpPlot(r.brain.rf, pch = 20, main = "")
```



As expected **GCS.15** (GCS=15 at 2h) is a strong predictor along with **bskullf** (basal skull fracture) and **age**. This means that most of the trees will have these predictors in the top split.

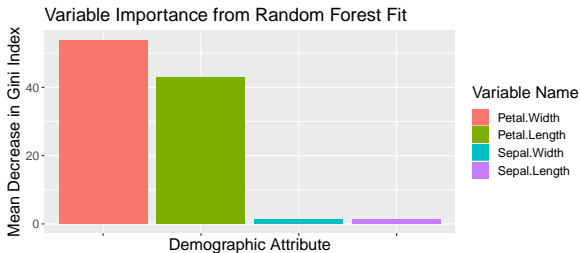
Iris example – bagging

(borrowed from <https://gist.github.com/ramhiser/6dec3067f087627a7a85>)

Variable importance plot for bagging

```
rf_out <- randomForest(Species ~ ., data = iris, mtry = 4, ntree = 500)
rf_out$confusion
```

```
##           setosa versicolor virginica class.error
## setosa         50          0          0         0.00
## versicolor     0          47          3         0.06
## virginica      0          3          47         0.06
```



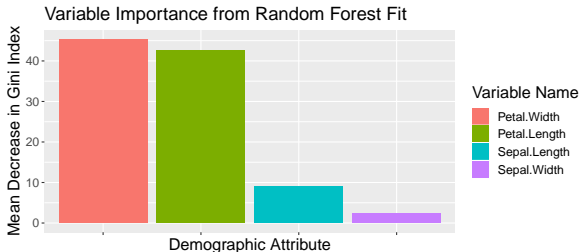
Iris example – random forest

(borrowed from <https://gist.github.com/ramhiser/6dec3067f087627a7a85>)

Variable importance plot for random forest:

```
rf_out = randomForest(Species ~ ., data = iris, mtry = 2, ntree = 500)
rf_out$confusion
```

##	setosa	versicolor	virginica	class.error
## setosa	50	0	0	0.00
## versicolor	0	47	3	0.06
## virginica	0	4	46	0.08



Boosting

Boosting is an alternative approach for improving the predictions resulting from a decision tree. We will only consider the description of boosting regression trees (and not classification trees) in this course.

In boosting the trees are grown *sequentially* so that each tree is grown using information from the previous tree.

- First build a decision tree with d splits (and $d + 1$ terminal nodes).
- Next, improve the model in areas where the model didn't perform well. This is done by fitting a decision tree to the *residuals of the model*. This procedure is called *learning slowly*.
- The first decision tree is then updated based on the residual tree, but with a weight.
- The procedure is repeated until some stopping criterion is reached. Each of the trees can be very small, with just a few terminal nodes (or just one split).

Algorithm 8.2: Boosting for regression trees

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \dots, B$, repeat:
 - 2.1 Fit a tree \hat{f}^b with d splits ($d + 1$ terminal nodes) to the training data.
 - 2.2 Update \hat{f} by adding in a shrunk version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$

- 2.3 Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$

3. The boosted model is $\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^b(x)$.

Boosting has three tuning parameters which need to be set (B , λ , d), and can be found using cross-validation.

Tuning parameters

- **Number of trees B .** Could be chosen using cross-validation. A too small value of B would imply that much information is unused (remember that boosting is a slow learner), whereas a too large value of B may lead to overfitting.
- **Shrinkage parameter λ .** Controls the rate at which boosting learns. λ scales the new information from the b -th tree, when added to the existing tree \hat{f} . Choosing a small value for λ ensures that the algorithm learns slowly, but will require a larger B . Typical values of λ is 0.1 or 0.01.
- **Interaction depth d :** The number of splits in each tree. This parameter controls the complexity of the boosted tree ensemble (the level of interaction between variables that we may estimate). By choosing $d = 1$ a tree stump will be fitted at each step and this gives an additive model.

Example: Boston data set

(ISLR book, Sections 8.3.2 to 8.3.4.)

Remember the data set: The aim is to predict the median value of owner occupied homes (in 1000\$)

We first run through trees, bagging and random forests - before arriving at boosting.

```
library(MASS)
set.seed(1)
train = sample(1:nrow(Boston), nrow(Boston)/2)
head(Boston)
```

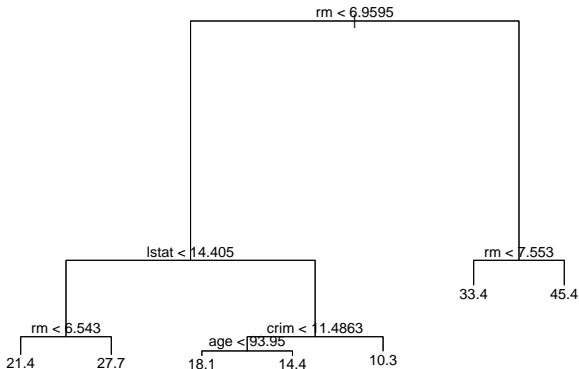
```
##      crim zn indus chas   nox   rm age   dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##      lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

Regression tree

```
tree.boston = tree(medv ~ ., Boston, subset = train)
summary(tree.boston)
```

```
##
## Regression tree:
## tree(formula = medv ~ ., data = Boston, subset = train)
## Variables actually used in tree construction:
## [1] "rm"      "lstat"   "crim"    "age"
## Number of terminal nodes: 7
## Residual mean deviance: 10.4 = 2550 / 246
## Distribution of residuals:
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -10.200  -1.780   -0.177    0.000   1.920   16.600
```

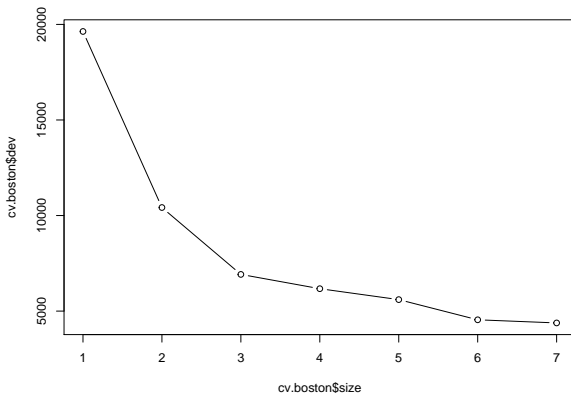
```
plot(tree.boston)
text(tree.boston, pretty = 0)
```



Remember: the `tree()` function has a built-in default stopping criterion to decide when to stop enlarging the tree. You can change this with the `control` option, for example by setting `control = tree.control(mincut = 2, minsize = 4, mindev = 0.001)`. You can then continue with the following steps from there.

Need to prune?

```
cv.boston = cv.tree(tree.boston)
plot(cv.boston$size, cv.boston$dev, type = "b")
```

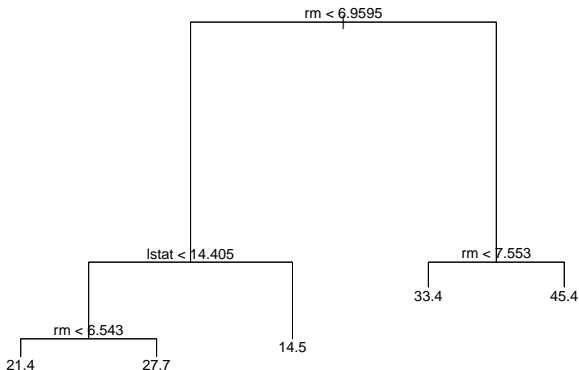


It looks like the most complex tree was selected.

Pruning

Want to prune anyway (even if most complex tree was selected)?

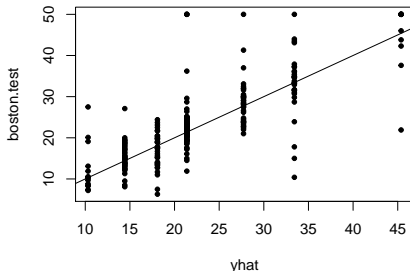
```
prune.boston = prune.tree(tree.boston, best = 5)  
plot(prune.boston)  
text(prune.boston, pretty = 0)
```



Test error for full tree

Because the full tree (with 7 nodes) was the best, we continue with this one.

```
yhat = predict(tree.boston, newdata = Boston[-train, ])  
boston.test = Boston[-train, "medv"]  
plot(yhat, boston.test, pch = 20)  
abline(0, 1)
```



```
mean((yhat - boston.test)^2)
```

```
## [1] 35.2869
```

Bagging

Remember: For bagging you can use the `randomForest()` function, but include all variables (here `mtry=13`).

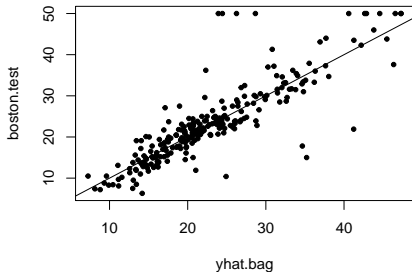
```
library(randomForest)
set.seed(1)
bag.boston = randomForest(medv ~ ., data = Boston, subset = train,
                           mtry = 13, importance = TRUE)
bag.boston
```

```
##
## Call:
## randomForest(formula = medv ~ ., data = Boston, mtry = 13, importance = TRUE,
##              Type of random forest: regression
##              Number of trees: 500
## No. of variables tried at each split: 13
##
##              Mean of squared residuals: 11.39601
##              % Var explained: 85.17
```

sub

Test error for bagged tree

```
yhat.bag = predict(bag.boston, newdata = Boston[-train, ])  
plot(yhat.bag, boston.test, pch = 20)  
abline(0, 1)
```



```
bag.boston = randomForest(medv ~ ., data = Boston, subset = train,  
  mtry = 13, ntree = 25)  
yhat.bag = predict(bag.boston, newdata = Boston[-train, ])  
mean((yhat.bag - boston.test)^2)
```

```
## [1] 23.66716
```


Random forest

Let's changing from bagging to a random forest (n.b., why are we now speaking of a *forst* and no longer of a *tree*?), using 6 randomly selected predictors for each tree:

```
set.seed(1)
rf.boston = randomForest(medv ~ ., data = Boston, subset = train,
  mtry = 6, importance = TRUE)
yhat.rf = predict(rf.boston, newdata = Boston[-train, ])
mean((yhat.rf - boston.test)^2)
```

```
## [1] 19.62021
```

It's interesting to see how the prediction error further decreased with respect to simple bagging.

Variable importance

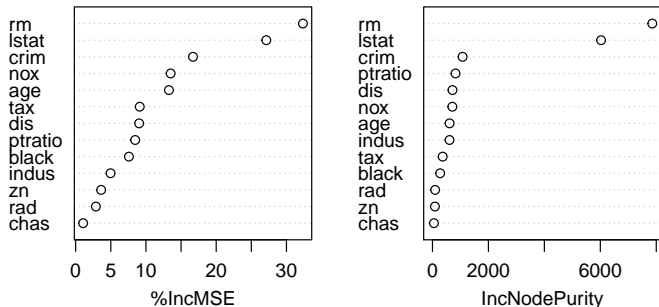
```
importance(rf.boston)
```

##		%IncMSE	IncNodePurity
##	crim	16.697017	1076.08786
##	zn	3.625784	88.35342
##	indus	4.968621	609.53356
##	chas	1.061432	52.21793
##	nox	13.518179	709.87339
##	rm	32.343305	7857.65451
##	age	13.272498	612.21424
##	dis	9.032477	714.94674
##	rad	2.878434	95.80598
##	tax	9.118801	364.92479
##	ptratio	8.467062	823.93341
##	black	7.579482	275.62272
##	lstat	27.129817	6027.63740

Interpretation?

And the variable importance plots

```
varImpPlot(rf.boston, main = "")
```



To understand what this means, please check again the meaning of the variables by typing `?Boston`.

Boosting

And finally, we are boosting the Boston trees!

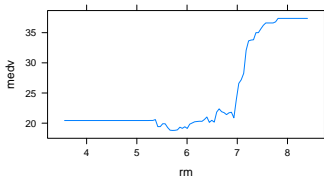
```
library(gbm)
set.seed(1)
boost.boston = gbm(medv ~ ., data = Boston[train, ], distribution = "gaussian",
  n.trees = 5000, interaction.depth = 4)
summary(boost.boston, plotit = FALSE)
```

```
##          var    rel.inf
## rm          rm 43.9919329
## lstat      lstat 33.1216941
## crim       crim  4.2604167
## dis        dis  4.0111090
## nox        nox  3.4353017
## black      black 2.8267554
## age        age  2.6113938
## ptratio    ptratio 2.5403035
## tax        tax  1.4565654
## indus      indus 0.8008740
## rad        rad  0.6546400
## zn         zn   0.1446149
## chas       chas 0.1443986
```

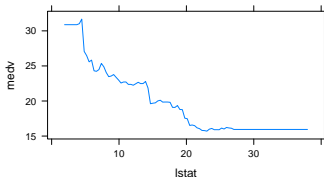
Partial dependency plots - integrating out other variables

rm (number of rooms) and **lstat** (% of lower status population) are the most important predictors. Partial dependency plots show the effect of individual predictors, integrated over the other predictors see Hastie, Tibshirani, and Friedman (2009), Section 10.13.2.

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



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



Prediction on test set

First for model with $\lambda = 0.001$ (default), then with $\lambda = 0.2$, to calculate the MSE on the test set. We could have done cross-validation to find the best λ over a grid.

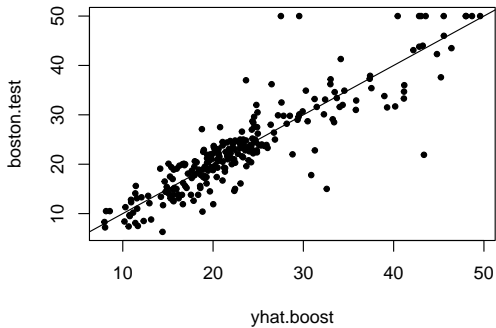
```
yhat.boost = predict(boost.boston, newdata = Boston[-train, ], n.trees = 5000)
mean((yhat.boost - boston.test)^2)
```

```
## [1] 18.84709
```

```
boost.boston = gbm(medv ~ ., data = Boston[train, ], distribution = "gaussian",
  n.trees = 5000, interaction.depth = 4, shrinkage = 0.2, verbose = F)
yhat.boost = predict(boost.boston, newdata = Boston[-train, ], n.trees = 5000)
mean((yhat.boost - boston.test)^2)
```

```
## [1] 18.33455
```

```
plot(yhat.boost, boston.test, pch = 20)  
abline(0, 1)
```



Summing up

with a Kahoot quiz or another online tool (SRS?)

Further reading

- Videos on YouTube by the authors of ISL, Chapter 8, and corresponding slides
- Solutions to exercises in the book, chapter 8

Acknowledgements

I thank Thea Roksvåg, who developed the set of slides that I modified to get this set of slides. In addition, Mette Langaas and Julia Debik have contributed to the original version.

References

- Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24: 123–40.
- Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference - Algorithms, Evidence, and Data Science*. Cambridge University Press.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Vol. 2. Springer series in statistics New York.
- Hyafil, L., and R.L. Rivest. 1976. "Constructing Optimal Binary Decision Trees Is Np-Complete." *Information Processing Letters* 5: 15–17.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.
- Ripley, Brian. 2019. *Tree: Classification and Regression Trees*. <https://CRAN.R-project.org/package=tree>.
- Ripley, Brian D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.