Module 8: Tree-based Methods

TMA4268 Statistical Learning V2021

Stefanie Muff, Department of Mathematical Sciences, NTNU

March 8 and 9, 2021

Example 2: Detection of Minor Head Injury

(Artificial data)

- Data from patients that enter hospital. The aim is to quickly assess whether a patient as a brain injury or not (binary outcome = classification problem).
- 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 personell that are not skilled in statistics, and
  - **fast**, such that the medical personell quickly can identify a patient that needs treatment.

$\rightarrow$ 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.

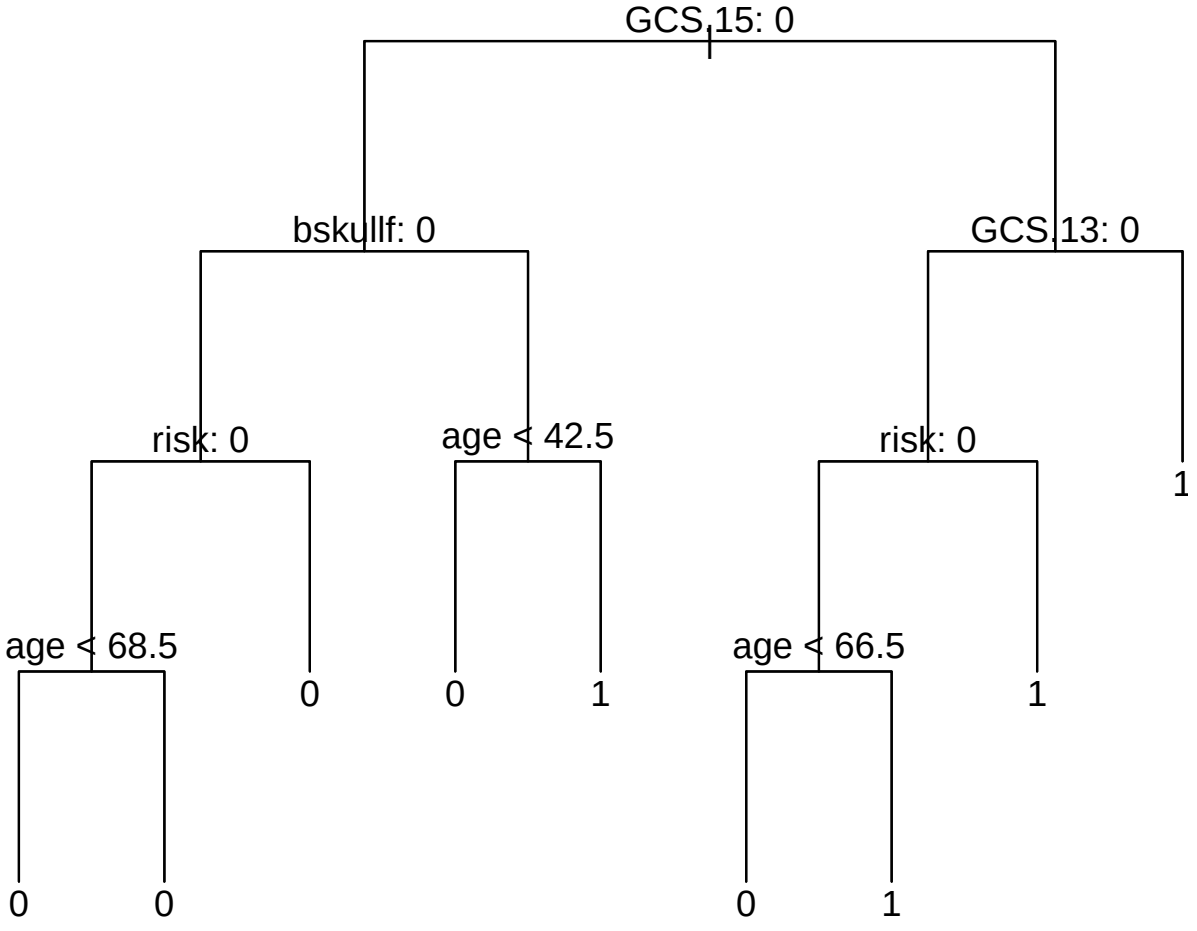
The dataset includes data about 1321 patients and is a modified and smaller version of the (simulated) dataset `headInjury` from the `DAAG` library.

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

- The variable `brain.injury` will be the response of our model (=1 if a person has an acute brain injury, =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) score[1] after 2 hours equal to 15 (or not)?
  - Has he/she an open scull fracture?
  - Has he/she had a loss of consciousness?
  - and so on.

[1] The 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.