Knowledge & the web: exercise session 6

Katrien Laenen   Gust Verbruggen   Ward Schodts

November 22, 2015

In this report we discuss which algorithms we tried to find a good classification model that predicts if a passenger of the Titanic will survive or not. To train the classification models, we use the information about passengers with passengerId 1 until 800 as our training data. The test data for evaluating each classification model consists of the passengers with ids 801 until 891.

1   Rule learning

One thing we tried was to learn classification rules from the training data. Rule learning algorithms come with different options for several aspects of the algorithms. In the exercise session we experimented with two of these. First, for choosing the conditions of a classification rule, there are two options: one is to add to the rule the condition which increases the accuracy the most, another is to add to the rule the condition which increases the information gain the most.

The first option resulted in a rule set consisting of 363 rules which clearly overfitted the training data. The second option resulted in this rule set:

- if Sex = male and Fare $\leq$ 15.173 then 0 (268 / 35)

- if Sex = female and Fare > 48.200 then 1 (4 / 76)

- if Sex = male and Cabin = ? then 0 (0 / 0)

- if Sex = male then 0 (152 / 62)

- else 1 (61 / 121)

which is correct for 617 out of 779 training examples. When evaluated on the test set, we got an accuracy of ?

A second aspect of rule learning is the desired pureness, which is the minimum ratio of the major class in a covered subset in order to consider the subset pure. The standard value for this is 0.9. We experimented with this value set to 0.8 and 1. For the desired pureness set to 0.8 and the information criterion we got the following rule set:

- if Sex = male then 0 (420 / 97)

- else 1 (64 / 189)

which is correct for 609 out of 770 training examples. When evaluated on the test set, we got an accuracy of ?. For the desired pureness set to 1 and the information gain criterion we got the following rule set:

- if Sex = male and Fare $\leq$ 15.173 then 0 (268 / 35)

- if Sex = female and Fare $>$ 48.200 then 1 (4 / 76)

- if Sex = male and Cabin = ? then 0 (0 / 0)

- if Sex = male then 0 (152 / 62)

- else 1 (61 / 121)

which is the same as the first rule set, which was obtained with the standard value for the desired pureness.

Next to these two, there are three other parameters for rule learning: sample ratio (which specifies the sample ratio of training data used for growing and pruning), minimal prune benefit (which specifies the minimum amount of benefit which must be exceeded over unpruned benefit in order to be pruned) and use local random seed (which indicates if a local random seed should be used for randomization). We did not experiment with these parameters, but kept the standard values for these, which were 0.9, 0.25 and no respectively.

From these results we can clearly see that Sex, Fare and SibSp are features which are important for this task. That Sex and Fare were important features was to be expected. When the Titanic collided with an iceberg, women and children and people of the higher ranks were evacuated first. This is reflected in the data as we see that women and people which payed a higher fare are more likely to have survived. That the number of siblings and spouses is also an important feature can be explained as follows. People in need want to stay close to their loved ones at all costs. But how larger a group, how more difficult it was to get on a lifeboat together. As such, how more siblings and spouses someone had aboard, how less likely that person was to survive.

2 Decision trees

Apart from *rule learning* we also experimented witch decision trees. First we applied the standard decision tree in RapidMiner, it created a to complicated decision tree, as you can see in figure 1.
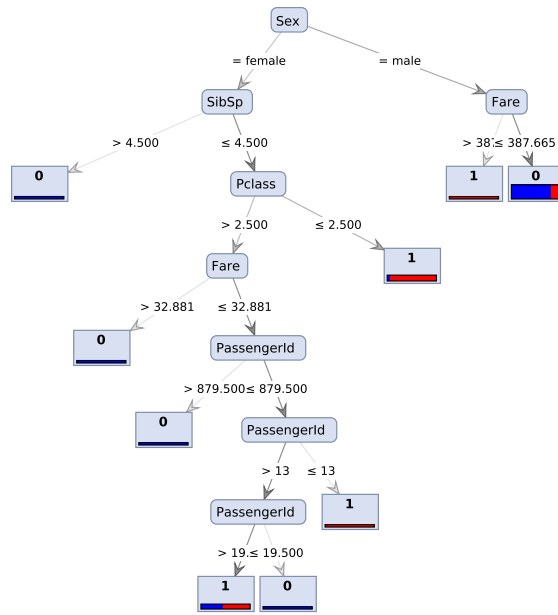

Figure 1: Original decision tree with maximal depth of 20

Next we limited the maximal depth of the decision tree to 3. This gave us figure 2, which is much more usefull.
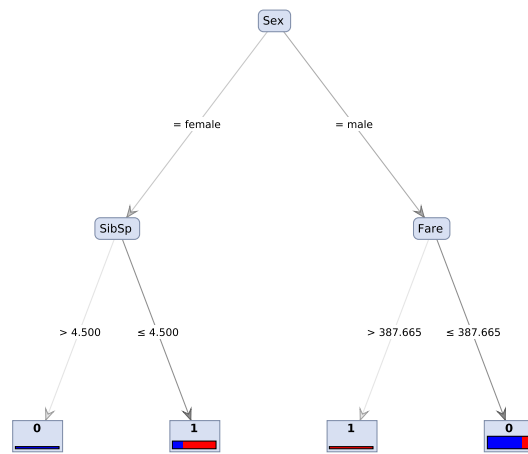
Figure 2: Decision tree with maximal depth of 3

3   ?