Machine Learning
Report 3

Sylwester Piątek

22 January 2019

Contents

1 Introduction

In this report I present the performance of models based on decision trees. The dataset used in this report contains data of passengers of the Titanic with the dependent variable **Survived** which tells if a passenger survived or not. The models shown in this report have different values of parameters such as **depth of the tree**, **maximum number of nodes** and **minimum number of samples** in one leaf. The basic model was built using an example from **kaggle.com**.

2 Preparing data

In order to allow the decision trees to use the data properly and to increase the accuracy of the models, some tranformation of the data had to be performed. The continuous variables like **Age** and **Fare** have been discretised (grouped into several bins). Variables **Pclass** (class which passenger used for traveling), **Sex** and **Embarked** (indicating city where a passenger entered the ship) have already been a discrete variable, but they have been transformed from strings to numbers. Moreover the variables **Parch** (which contains number of parents or children of the passenger) and **SibSp** (which contains number of siblings or spouses of the passenger) have been changed into binary variables meaning if someone had parents/children or not and if someone had siblings/spouses or not. The missing values in columns **Age** and **Fare** have been replaced with median values. The ones in **Embarked** have been replaced with the most frequent value which was **C**.

3 Different models

I verified how does changing certain parameters of the model influence on accuracy of the model

3.1 Basic model

Basic model which had default values of parameters, without pruning, achieved accuracy 0.9027, which is quite a good result regarding this dataset. The visualisation of the model is available in a file *titanic_ basic_ model.pdf*.

3.2 Basic model with cross-validation

The model with 5-fold CV had the best accuracy. The difference between 5-fold and 10-fold was small and could be random.

n	ACC
3	0.7800
5	0.8093
10	0.8059

Table 1: Accuracy of the models with various number of cross-validation iterations

3.3 Comparing different criteria of choosing variables

The default criterion is the one using *Gini impurity*. Replacing it with*entropy impurity* did not change the accuracy of the model. Checking the details of the models allows us to realise, that the model did not change. We have to remember though, that in different dataset choosing a proper criterion would matter.

3.4 Changing maximum depth of the tree

The default model has no boundaries of how the tree is growing. I compared the accuracy of the trees with depths 4, 7 and 10. The last one is very similar to the basic model. The visualisation of the tree with depth 4 is available in a file *titanic_ max_ depth_ 4.pdf*.

n	ACC
4	0.8473
7	0.8713
10	0.8997

Table 2: Accuracy of the models with different depth of the decision tree

3.5 Changing maximum number of the leaves

The accuracy of the model with up to 10 leaves is $ACC = 0,8503$ which is a good result for such a simple model. The trees with limitations of 16 and 23 leaves had slightly better score. The model with up to 100 leaves had almost as good accuracy (0.9012) as the basic model (0.9027). Pictures of the trees with the values of these parameters equal to 10 and 33 are available in the files *titanic_ notes_ 10.pdf* and *titanic_ notes_ 33.pdf*.

n	ACC
10	0.8503
16	0.8563
23	0.8653
33	0.8743
50	0.8892
100	0.9012

Table 3: Accuracy of the models with different maximum numbers of leaves of the decision tree

3.6 Changing the strategy of selecting variables

Changing the parameter *splitter* in a function building the model from the value *best* (which takes the best possible variable) to *random* (which makes the model choose the variables randomly) does not change the accuracy of the basic model, but it can decrease the accuracy of pruned trees. For example it decreased the accuracy of the model with maximum number of leaves = 33 from 0.8743 to 0.8698.

3.7 Changing the minimum number of samples in one leaf

The default value of this parameter is 1. Increasing this value decreases the accuracy and the deifference is noticeable. Comparison of the accuracy for numbers of samples $n \in \{1, 2, 3\}$ is available in a table below. Picture of the tree with $n = 3$ is available in *titanic_ min_ sample_ 3.pdf*.

n	ACC
1	0.9027
2	0.8802
3	0.8668

Table 4: Accuracy of the models with different depth of the decision tree

4 Summary

The basic model has an accuracy slightly higher than 90%. There is though a risk, that it is overfitted. Smaller trees (pruned with various parameters) achieve accuracy higher than 85%.