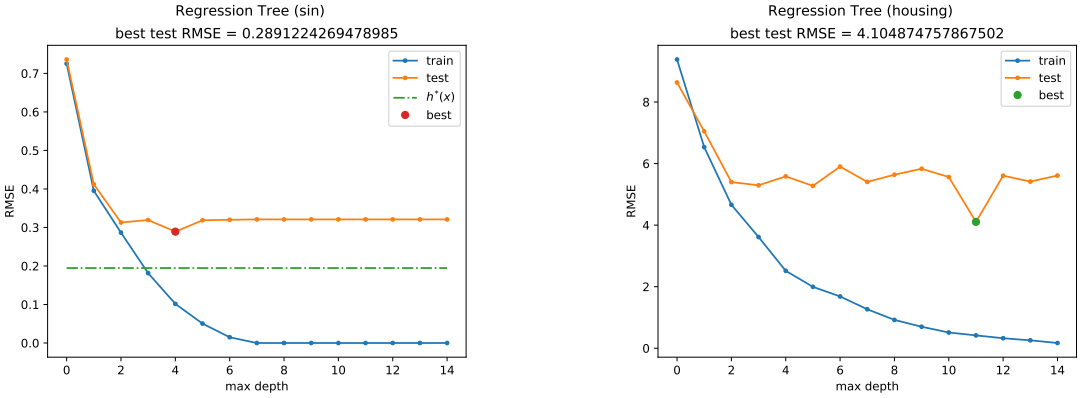
# SSU 05 – Ensembling

Lukáš Hromadník

## 1 Assignment 1




(a) Sin dataset

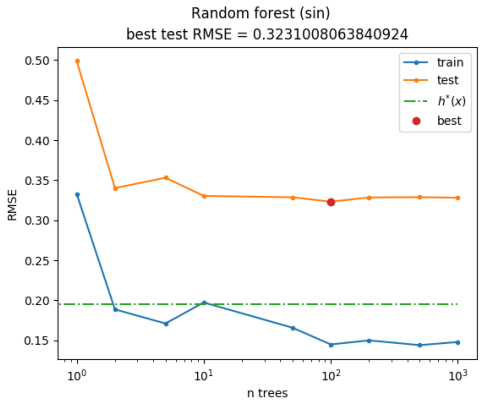
(b) Boston housing dataset

Obrázek 1: Assignment 1 results

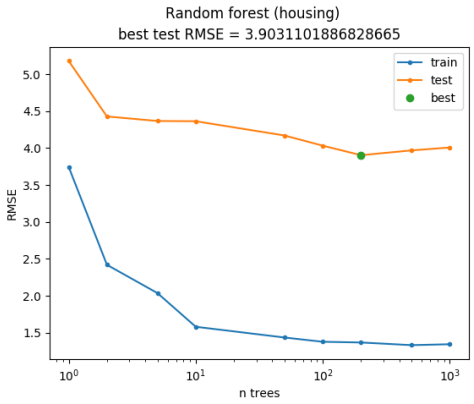
The figure 1a in which sin dataset is plotted shows that the optimal depth for the tree to grow is 3. From the depth of 7 and more the tree overfits training data and cannot improve itself. Overfitting is visible on the test data too. The error on the test data is the same from a depth of 7.

The Boston housing dataset in figure 1b has more parameters to learn. In the result we can see that the optimal depth has value of 11. The maximal depth in this example was not enough to overfit the given data. There is a noticaeble trend that in a deeper tree the classifier will overfit the data.

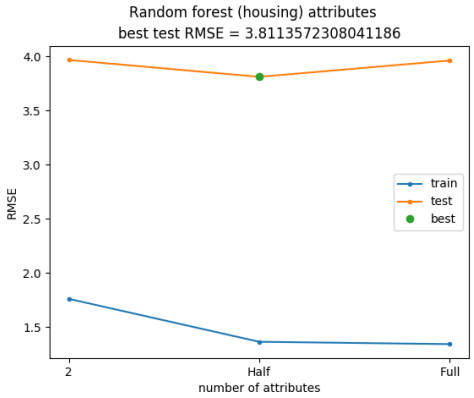
# 2 Assignment 2




(a) Sin dataset




(b) Boston housing dataset




(c) Number of attributes used for learning

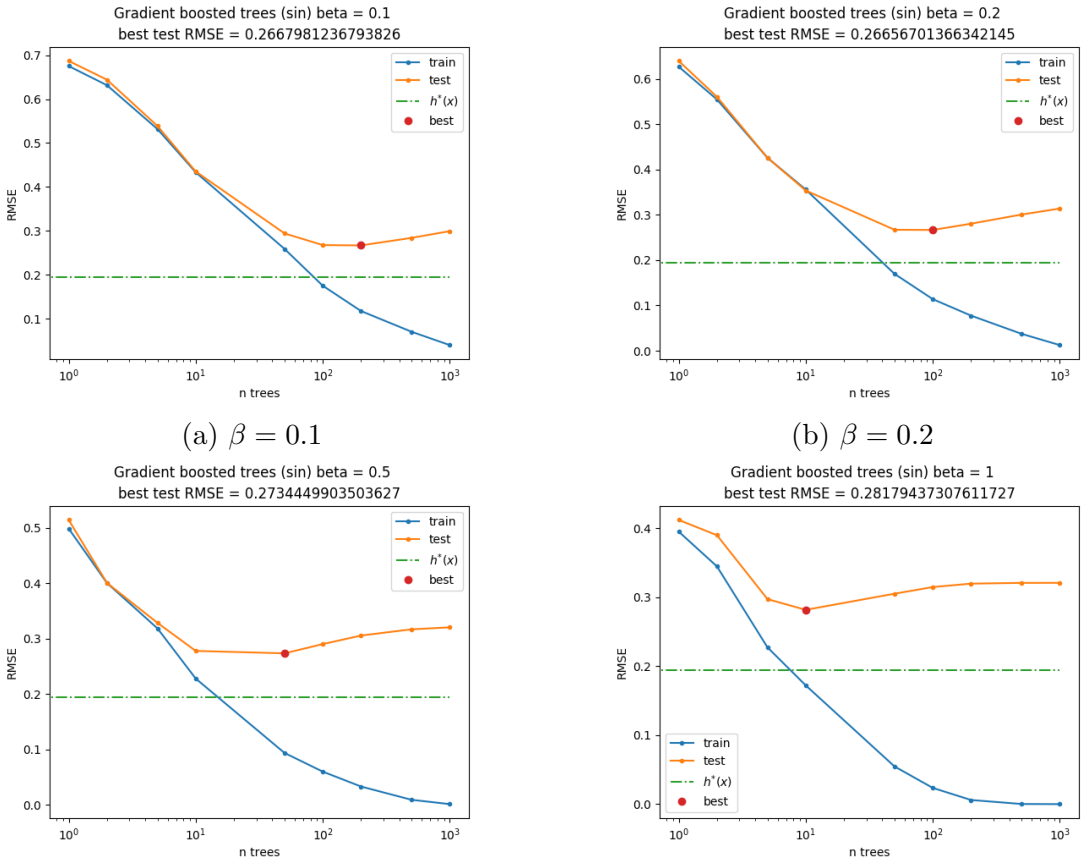
Obrázek 2: Assignment 2 results

From the first figure 2a it is clear that optimal number of trees in Random forest learned on a sin dataset is 100. Another interesting thing is that adding more trees to the forest doesn't improve final result.

Second figure 2b shows housing dataset. The optimal number of trees for this dataset is 200. Same thing as in the previous result appeared here that more trees don't improve the final result.

Last figure 2c shows performance given a number of attributes to learn in the tree. Best result is given by using half of the attributes from the dataset.

# 3 Assignment 3




(a) $\beta = 0.1$

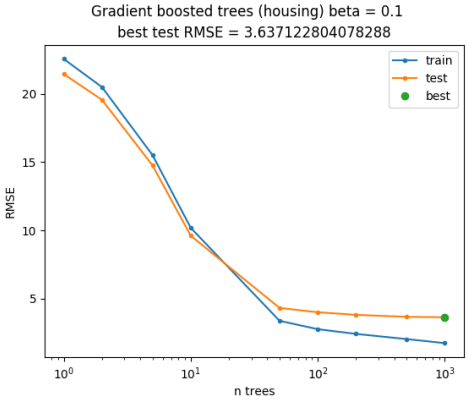
(b) $\beta = 0.2$

(c) $\beta = 0.5$

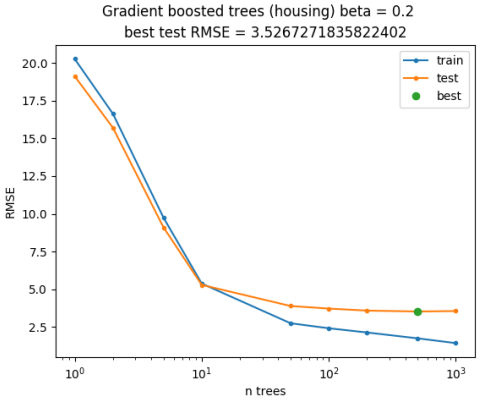
(d) $\beta = 1$

Obrázek 3: Assignment 3 Sin dataset results

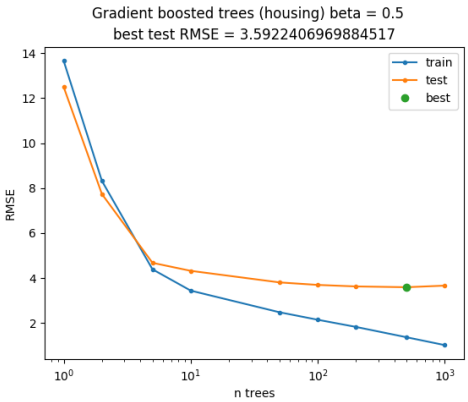
There are four graphs in 3 depicting learning of Gradient boosted trees with different parameter $\beta$ given sin dataset. The best result is given by $\beta = 0.2$ and it's RMSE $= 0.2666$. Another thing that can be seen from the graphs is that with increasing value of $\beta$ the number of trees needed for the best value is decreasing.



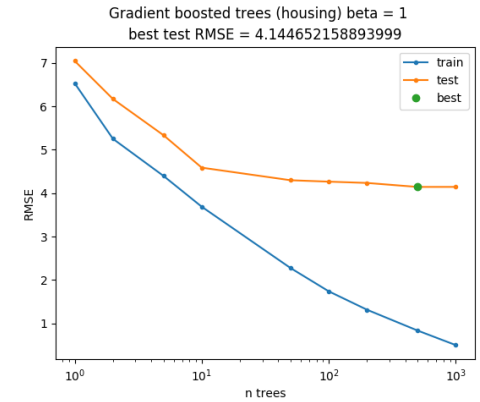

(a) $\beta = 0.1$




(b) $\beta = 0.2$




(c) $\beta = 0.5$




(d) $\beta = 1$

Obrázek 4: Assignment 3 Housing dataset results

Gradient boosted trees results given housing dataset in figure 4 are more or less the same. The best result is given by $\beta = 0.2$ and it's RMSE $= 3.5267$. Here with the increasing value of $\beta$ the number of trees used in for the best value doesn't decrease. That's probably because the dataset is more complex than the previous one.

# 4 Assignment 4

| Number of threads | Duration time |
|---|---|
| 1 | 500.25 s |
| 4 | 148.00 s |
| 6 | 142.65 s |
| 8 | 143.83 s |

The best result is given with 6 threads. This result was measured on the 4-core / 8 threads system. Using the same number of threads for computation as the maximal number of threads for the CPU isn't the best choice because there is always at least one thread (the main thread) which is used by the system. That's the reason why in my case it was the best to use 6 threads.

The implementation is available inside the `RandomForest` class.