Report Machine Learning 07.11.2019
Lab 1

Laurenz Ohnemüller
i6203321

Train one-level decision trees and multi-level decision trees on the two data sets. Determine the accuracy rates of the resulting classifiers using the training set and 10-fold cross validation.

- Is there a difference?

  BREAST CANCER:

  Our average cross-validation score for the 'One Level Decision Tree' is: 0.909
  Our average cross-validation score for the 'Multi Level Decision Tree' is: 0.941

  Here, we can see that the score for the multi-level decision tree is a bit higher than for the one-leve decision tree, which means that the mutli level decision tree is slightly more accurate in this case.

  DIABETES:

  Our average cross-validation score for the 'One Level Decision Tree' is: 0.715
  Our average cross-validation score for the 'Multi Level Decision Tree' is: 0.689

  Here, we can see that the score for the muli-level decision tree is a bit lower than for the one level decision tree, which means that it performed a bit worse than the one-level decision tree.

- Can you explain why?

  BREAST CANCER:

  Normally, it would make sense that the multi-level decision tree (dt) performs better since we can capture more factors with it. To achieve a good prediction we want to find the 'sweet spot' between underfitting and overfitting, where we move from underfitting to overfitting the more leaves we allow. It just may be the case that the one-level dt suffers a bit from underfitting and that therefore the mutli-level dt is more accurate.

  DIABETES:

  Here, the exact opposite is the case. The multilevel dt suffers under overfitting which makes the one-level dt more accurate.

- Analyze the resulting classifiers from a comprehensibility point of view.

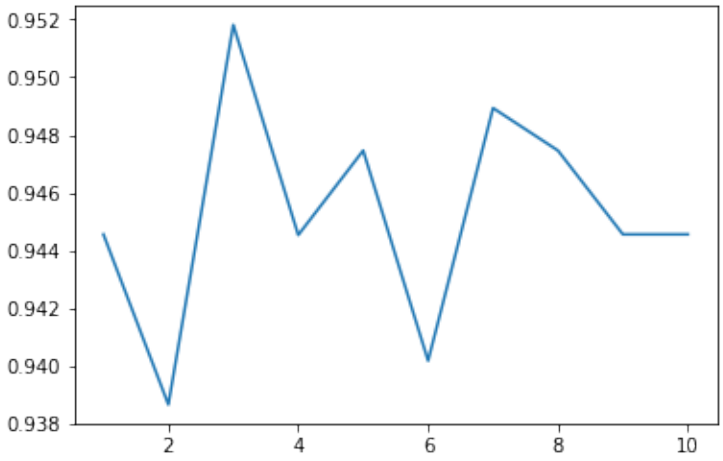
  We create two seperate decision trees, one with just one-level which equals one split, and another one with multiple levels (multiple splits). On the one hand, the more splits we have, the better predicitons to the actual values we can make. However, since this makes each leave with only a few elements it can become very distinct. Now if we have to many splits our model can suffer from overfitting. On the other hand, when we have just one-level we might not cover important patterns and features which leads to underfitting. Both underfitting and overfitting lead to to less acurate predictions. Therefore, we want to find the perfect spot between these two.

Experiment with multi-level decision trees and error pre-pruning by changing the option min_samples_leaf from 0 to the size of the datasets (use some step). Estimate the accuracy rates of the resulting decision trees using the training set and 10-fold cross validation.

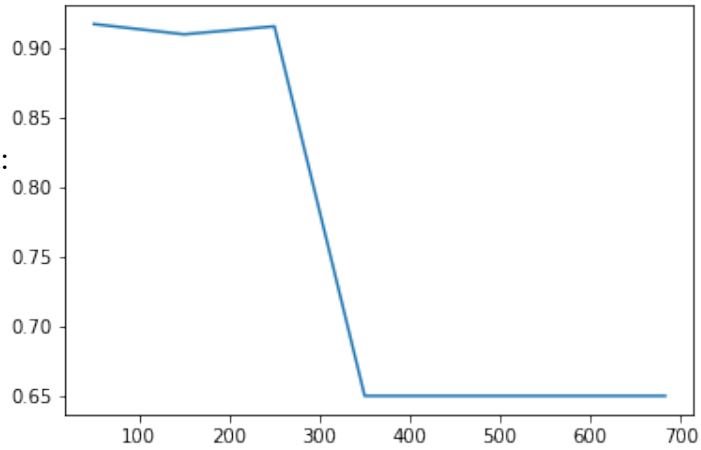
- Plot the 2 accuracy rates based on the training set and 10-fold cross validation for min_samples_leaf from 0 to the size of the datasets. Identify the regions of underfitting, optimality, and overfitting. Explain how you have identified these regions.

  Here are the plots for breast cancer:
  (note: the x-axsis describes the depth of the tree and the y-axsis is the cross-validation score)

  For the step size from 1 to 10:



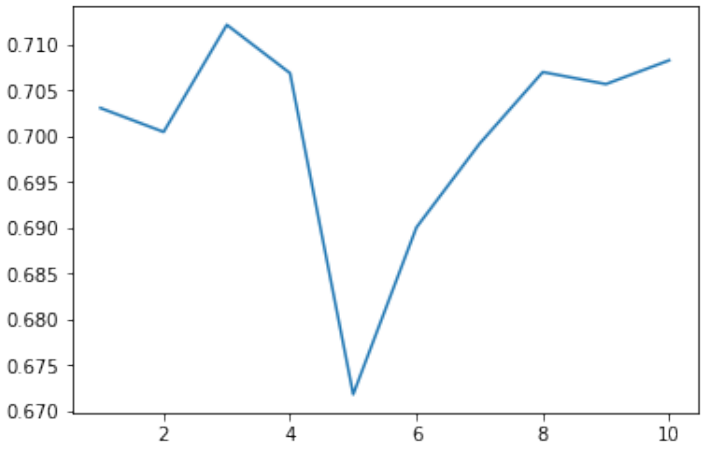
  For the step size from 50 to the end:



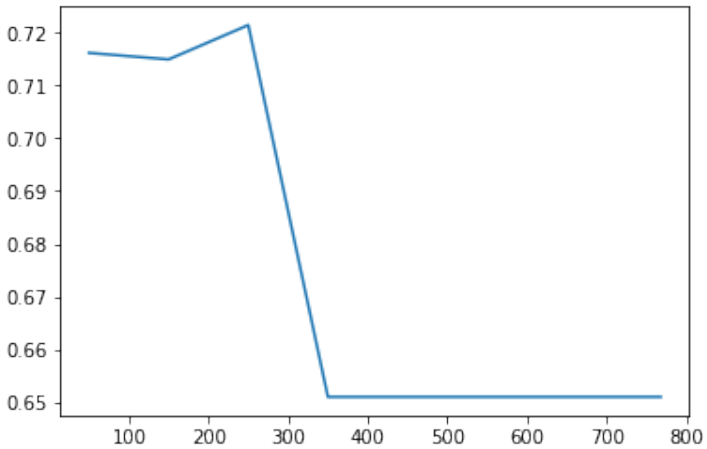
In generel, we can see that for the breast cancer data the cv-score is lowest for a higher depth of the tree. Interesting to see is in the first graph that we almost have no underfitting, we are in a region where the change is ‘only’ about 0.02. To that, we can see that we reach the highest score for just the step size 2. In graph 2, we can see pretty good that the region from 50-300 is an optimal region (including the 1-300 - its almost a straight line), while our model suffers from overfitting at about 300, which gets worse and worse. From 350 to the end the accuarcy does not change anymore. We can identify in this case almost no underfitting.

Here are the plots for diabetes:

For the step size from 1-10:



For the step size from 50 to the end:



For the diabetes data, we can make interesting observations: from the first graph and the second, we can clearly see our model suffering form underfitting until the stepsize of abour 150. At about 300 we can identify our optimal region, where the model is the most accurate. After that the model quickly begins to suffer from overfitting.