Machine Learning Worksheet 2

Tomas Ladek
3602673
tom.ladek@tum.de

Problem 1

After parsing the data that was given in a csv file and making up an efficient data structure (matrix), the Gini index of the root node ($C = \{0, 1, 2\}$) was calculated:

$$i_G(t) = 1 - (\frac{5}{15})^2 - (\frac{6}{15})^2 - (\frac{4}{15})^2$$

Then in 0.1 steps from -0.6 to +10.0 (limits determined by data inspection), the Gini index of all possible left/right splits of the root node was calculated, for all features $x_{i,1}...x_{i,3}$. The maximum was a difference in Gini indices of ≈ 0.3615 for the first feature ($x_{i,1}$) for a split at the value 4.5[1]. The formula used for calculating the difference was

$$\Delta i_G(t)(x_{i,1} \leq s, t) = i_G(t) - \frac{\#classes\ in\ left\ tree}{\#classes\ in\ current\ data\ set} i_G(t_L) - \frac{\#classes\ in\ right\ tree}{\#classes\ in\ current\ data\ set} i_G(t_R)$$

with i_G being the Gini indices of the current node, the left tree and the right tree respectively.

Splitting the root node at $x_{i,1} = 4.5$ yielded a left tree (values less than or equal to the split value) consisting of a pure node (class distribution of 100% for class '1', Gini index 0) and a right tree combining the remaining classes '0' and '2', Gini index 0.4938. No further splits in the left tree were needed.

Once again performing a maximalization on the Gini index differences for every possible split of every feature in 0.1 steps yielded feature $x_{i,1}$ at 7.4 as the best split. The result was a left sub-tree with class distribution '2': $\frac{2}{3}$; '0': $\frac{1}{3}$ and a right sub-tree with a pure distribution of class '0'. The corresponding Gini indices were 0.4444 and 0 respectively.

As the maximum requested depth was reached, no further splitting was considered. The (raw) python code can be provided on request.

[1] Due to the chosen calculation procedure, the split values take the maximum possible value between two different splits, instead of the average.