Saturday, December 3, 2016

Decision tree with pruning

Preprocess

In the preprocess step, we processed on continuous typed features so that they can be handled in the following tree node splitting step. We discretized each continuous attribute to form an ordinal binary attribute by checking if a sample is smaller or bigger than the median of all samples of the attribute. This way, continuous attributes were labeled either 0(no bigger than the median of the attribute) or 1(bigger than the median).

For nominal attributes, we numbered them to make implementation easier. For dataset2, “Absent” was labeled as 0 and “Present” was labeled as 1.

Tree splitting criterion

We used 2-way splitting to split tree nodes and we used classification error to choose the best split.

choose_i =

arg min{ p_i*classification_error_i(left_child) + (1-p_i)*classification_error_i(right_child) }

If a node doesn’t meet the stoping criterion(not leaf node), we followed the splitting criterion to split.

Stoping criterion: check if a node is leaf node

1. has only one point in the node
2. all points in the node have same label(impurity is 0)
3. all splits’s impurity is bad: wighted sum of children > parent
4. all points in the node have same feature situation(line 119)