4.1.4 Evaluation Measures

# Original

## Classification Performance

For measuring performance, we use a weighted f1-measure where the inputs are the predicted labels and the black-box labels (rather than the real class labels). This was chosen as we assume each class is equally important, and wish to have a valid measure for both binary and multi-class classification. This measure can be roughly interpreted as "how well are we able to reconstruct the predictions of the black-box classifier for each class". For presentation, we scale this to the range 0 . . . 100, rather than 0 . . . 1.

## Complexity

Measuring complexity across classifiers can be a complex task, however, here thankfully since each of the comparison methods is somewhat similar in representation, there is a natural definition of complexity. We define complexity as the number of splitting points in a tree, where in the proposed method, if a constructed feature is used as a split, this counts as multiple splits, i.e. f1+f1<=0, would be a complexity of 2, rather than 1, to provide a fair comparison. For Bayesian rule lists, the complexity is the number of rules + the number of conjunctions in these rules, i.e. if f 1 = 2 ∧ f 2 = 0 then ... counts as 2. The number of rules in logistic regression is measured as the number of non-zero coefficients. Therefore, for all methods, the minimum complexity is 0 (i.e. predict majority class, no rules learnt), and the maximum approaches ∞.

# Condensed

## Classification Performance

For measuring performance, we use a weighted f1-measure. For presentation we scale this to the range 0…100.

## Complexity

We define complexity as the number of splitting points in the tree. If constructed features are used they count as multiple splits (f1+f2<=0 would count as 2). For Bayesian rule lists the complexity is measured as the number of rules + the number conjunctions in these rules. For logistic regression complexity is measured as the number of non-zero coefficients.