2.1 Model Extraction

# Original

One of the early works in the area was [7] which used decision trees to approximate complex black-box models. Similar work was done in [1]. Both utilise decision trees as the simple method for approximating more complex models. However, an issue is ensuring these decision trees remain simple themselves, so operations such as early stopping or pruning become essential. Furthermore, due to the greedy construction of the trees, they may not be the best approximator of the more complex black-box methods. Other methods such as logistic regression can also be used, and these are compared in Section 4.

Bayesian Rule lists [20, 35] are an approach to IML which aim to achieve a good balance between complexity and accuracy, by using Bayesian optimisation to generate a set of "if...then..." statements which can be used for prediction. The idea is that these resulting rules are simple and easily interpretable.

Model compression was proposed in [5] (and expanded in [17]), where the authors use a neural network to compress large complex ensembles (often with thousands of base members) by training the smaller neural networks on the predictions of the ensemble. While these are not directly related to interpretable machine learning (as the neural network learnt is not necessarily interpretable), the concept is similar, and these works showed the simpler model can often achieve similar error rates to the larger, more complex ensembles, so this is promising for model extraction methods.

# Condensed