Online Supplementary Material.

An Explainable AI Handbook for Psychologists: Methods, Opportunities, and Challenges

Lavelle-Hill, Rosa Smith, Gavin Deininger, Hannah Murayama, Kou

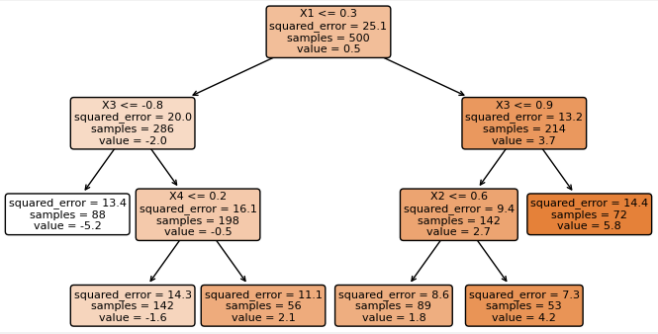
Author note: Correspondence concerning this article should be addressed to Rosa Lavelle-Hill, University of Basel, Email: rosa.lavelle-hill@unibas.ch.

Decision Trees

Decision trees are a type of supervised machine learning algorithm that can be used for both *classification* (where the outcome is a categorical variable) and *regression* tasks (the outcome is continuous). They can handle predictor variables with different data types, as well as missing data (Breiman et al., 1984). A tree is fit to data by using (often binary[a]) decisions on a predictor variable to partition the data into smaller and smaller subsets (referred to as 'nodes') so that the homogeneity (or purity) of the outcome variable is maximized within each new subset. The prediction for a certain data instance is the mean or majority class of the final subset that the data instance ends up in. The split decision process involves evaluating different split options for each feature to find the split that best maximizes the between-group variance and minimizes the within-group variance (or purity for classification). For further reading on decision trees and their extensions (e.g., random forests (Breiman, 2001) and extreme gradient boosting (Chen and Guestrin, 2016)), we recommend Costa and Pedreira (2023); Rokach (2016).

Decision trees are useful for producing explanations, as they are inherently interpretable models. The tree-like structure, similar to a family tree diagram, can be drawn or printed (see below). The branches (lines) represent the different split pathways, and the terminal (or 'leaf') nodes contain the data used to make the final prediction. For each data point in a leaf node, one can follow the decision pathway down the tree that leads to that data instance's prediction, producing local explanations. Decision trees can additionally be used to detect non-linear interaction or threshold effects by analyzing the sequence of decisions and the cut-off points used. They are also intuitive to understand as they mimic how real-world decisions are made (i.e., medical diagnostic decisions). For these reasons, decision trees are also a useful option for global surrogate models — an interpretable model that is trained to approximate the predictions of a more complex "black box" model. In this case, researchers first train a black box model to make an accurate prediction about the outcome. After obtaining the best model, the decision tree is trained with the same set of inputs to predict the black box model's predictions rather than the outcome variable (Molnar, 2022). The resultant decision tree should approximate how the black box model makes predictions in an interpretable manner.

However, decision tree models also have their limitations (Rokach, 2016). First, they are highly unstable. Often, a small change in data can result in a highly variant decision tree structure (Li and Belford, 2002; Breiman et al., 1984; Quinlan, rint). To try and overcome this limitation, random forests can be used to produce powerful predictions using many decision trees (Breiman, 2001). However, this comes with the trade-off of reduced interpretability. Second, it is important to consider when interpreting their structure that they are "greedy" models. This means that they only consider optimizing the immediate split decision. The algorithm does not "look ahead" to find the most optimal global set of decision pathways (but see Zhu et al., 2015, for an extension using reinforcement learning to overcome this). In such cases, practitioners may end up with suboptimal models with explanations that differ from what would be revealed under an optimal model.



This tree was created with a maximum depth of three and at least fifty samples in each leaf node. Within each node, first, you can find the split decision (for the non-terminal nodes), for example, X_1 <=0.3. If the statement is true, the data goes to the left branch and false to the right. The *squared error* is the measure of impurity, where less is better. *Samples* refer to the number of data instances in the node, and the *value* is the prediction at that node (the mean value).

[a] Although multiway splits can be implemented, they tend to break up the data too quickly, leaving insufficient data for the daughter nodes (Hastie et al., 2009).

Artificial Neural Networks

Artificial Neural Networks (ANN) are a family of powerful machine learning algorithms inspired by how neurons work in the brain. Their development is partly rooted in psychology (e.g., Rumelhart et al., 1986). ANNs consist of layers of nodes and directed edges (connections). The first layer is an input layer, the last layer is the output layer, and in between lies hidden layer(s). At each node, input values get converted to output values using the linear combination of inputs, a bias term, and a non-linear activation function (Haykin, 1998). The weights can be conceptualized similarly to regression coefficients; and the bias term, the intercept. The activation function differs for different types of neural network architecture, but a simple example is a sigmoid function used in logistic regression. Other common activation functions include the tangent hyperbolic (tanh) function and the rectified linear activation function (ReLU) (Hastie et al., 2009). The network learns the weights and bias terms by first randomly setting them (or setting the bias terms to zero) and then passing data forward through the network. The information on the difference between the model's prediction and the data's actual output value (the error) is then propagated back through the network (for more information on back-propagation see Hecht-Nielsen, 1992; Werbos, 1990) to update the weights in a manner that minimizes the loss function used for the error, for example, using a (stochastic) gradient descent algorithm (Hastie et al., 2009). This process is known as one epoch and is repeated for each instance of data to train the network by updating the connection weights and bias terms. In the ANN family, there are different types of network designs, network structures, and activation functions suited to different tasks and input data (for more information on neural networks, see Bishop, 1995; Haykin, 1998; Hastie et al., 2009).

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York, NY, USA. 10.1093/oso/9780198538493.001.0001.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. 10.1023/A:1010933404324.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. wadsworth int. *Group*, 37(15):237–251.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Costa, V. G. and Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 56(5):4765–4800.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Haykin, S. (1998). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier.

Li, R.-H. and Belford, G. G. (2002). Instability of decision tree classification algorithms. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 570–575.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Self-published via Christoph Molnar, 2nd edition.

Quinlan, J. R. (2014 (reprint)). *C4.5: Programs for Machine Learning*. Elsevier (originally 1993), 4th edition. Reprint of the classic machine learning monograph.

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, 27:111–125. 10.1016/j.inffus.2015.06.005.

Rumelhart, D. E., Hinton, G. E., McClelland, J. L., et al. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(45-76):26.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Zhu, R., Zeng, D., and Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784.