Group 4: Balija Sree B., Elavarthi Pragna, Parupudi, V.S.Raghu R.K., Sangal Rupashi

Problem: With the advent of advanced AI algorithms like transformers, neural networks etc. researchers across the globe are able to find solutions to problems considered unsolvable 20 years ago. However, many, if not all, of theses models are black boxes. They solve the problem efficiently once trained well, but do not help researchers make inferences. Hence, their scope is limited to fields where interpretability is not very important. For example, in speech recognition and personal voice assistants, we are not concerned about how the neural nets are working, as long as they do.

However, there are important fields where inference is everything, like policy making in economics or in medical research. A medical researcher needs to know if high blood pressure is primarily driven by higher sodium intake or the lack of physical exercise (or both). So that the medical professionals can make an educated recommendation.

A new and novel development in this field is "interpretable AI" where we utilise the power of black box models, but also constrain them to get an interpretable model. The accuracy of such models may be a little lower than unconstrained black boxes, but as discussed above, we have fields where interpretability is more important. In this exercise, we use this idea to build our own interpretable neural network to a problem in the medical domain.

Data: We use the Framingham dataset (Dileep, 2018). The data is from a cardiovascular study on residents of Framingham, Massachusetts. The goal is to predict whether a subject will have coronary heart disease (CHD) 10 years into the future. The dataset has 4000 rows and 15 features, having demographic and clinical attributes.

Proposed Solution: A good implementation of the idea mentioned above is using Neural Additive Models (NAMs) (Agarwal, 2021). In these models, each feature is input to a neural network (only one feature is input to each neural network). The outputs of these neural nets are combined with learned weights to get the final target. Note that, except at the output, no two features are ever combined. This helps keep the model interpretable. Mathematically, this is the same as doing an independent kernel transformation on each variable and then running a simple linear regression. The power of the model comes from the neural networks (each neural network does a kernel transformation), the interpretability comes from the "purity" of features i.e. each feature is a kernel transform of the initial input feature.

Experiments: Though the idea is very interesting, we need to do a good number of experiments to validate it. The gold standard of interpretability is the linear regression (with all the BLUE assumptions), so, we shall run a linear regression to get the baseline AUC or MSE (regression is the least powerful so we take it as a baseline). Then the plan is to implement a few algorithms of increasing complexity (RF, XGBoost, fully connected neural networks). We want to see the accuracy of NAMs compared to the above models. Ideally, we want an accuracy closer to the fully connected neural net while being interpretable. Also as the problem is that of classification, we can use AUC as a metric to evaluate our model. Interpretability can be evaluated by using existing literature as a reference.

References:

[1] Dileep. (2018, June). Logistic regression To predict heart disease, Version 1. Retrieved February 3, 2023 from https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression.

[2] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, Geoffrey E. Hinton, Neural Additive Models: Interpretable Machine Learning with Neural Nets, Advances in Neural Information Processing Systems 34 (NeurIPS 2021).