Interpretable Neural Networks

Interpreting black box models is a significant challenge in machine learning, and can significantly reduce barriers to adoption of the technology.

There are two things to note from this ice cream model: firstly, the effect of the features is being compared to a baseline of what the model would predict when it can't see the features. Secondly, the sum of the feature importances (the red and blue arrows) is the difference between this baseline and the model's actual prediction for Bob.

Unfortunately, while certain machine learning algorithms (such as XGBoost) can handle null feature values (i.e. not seeing a feature), neural networks can't, so a slightly different approach will be needed to interpret them. The most common approach so far has been to consider the gradients of the inputs with respect to the predictions.

In this post, I will cover the intuition behind using these gradients, as well as two specific techniques that have come out of this: Integrated Gradients and DeepLift.

Using gradients to interpret neural networks

Possibly the most intepretable model — and therefore the one we will use as inspiration — is a regression. In a regression, each feature x is

assigned some weight, w, which directly tells me that feature's importance to the model.

Specifically, for the ith feature of a specific data point, the feature's contribution to the model output is

$$w_i \times x_i$$

What does this weight w represent? Well, since a regression is

$$Y = (w_1x_1 + w_2x_2 + \ldots + w_ix_i + \ldots + w_nx_n) + b$$

Then

$$w_i = \frac{\partial Y}{\partial x_i}$$

In other words, the weight assigned to the ith feature tells us the gradient of that feature with respect to the model's prediction: how the model's prediction changes as the feature changes.

Conveniently, this gradient is easy to calculate for neural networks. So, in the same way that for a regression, a feature's contribution is

$$w_i \times x_i = x_i \times \frac{\partial Y}{\partial x_i}$$

, perhaps the gradient can be used to explain the output of a neural network.