October 1986

Learning Representations by back-propigating errors.

Reading and reviewing this paper in 2024, 38 Years Later gives you a diffrent Perspective that if you were to read this back in 1986.

The authors state that they will introduce a new method of learning. They will be minimizing Diffrance between to output & the desired output. To do this they will be adjusting weights in these "hidden layers" and these hidden layers will learn new feature & relationships that were not part of the input or output.

They note that many attemps have been made do find an update rule that can turn an abritrarly conected network into one that can perform well on a certin task. Now they say this is relativly simple if input, output are directly connected. But get's harder if we have these hidden units in the middle. These hidden units must both turn on and off meaning they must learn what to reprisent. They claim to have a simple procidure to build these internal representations.

They then go on to descripe what we would know as an MLP. The mul + sum and also use sigmoid as a non-linearity but state that other functions can be used as long as they are differentiable.

They then go on to define how the loss is computed to where it is a MSE loss function.

Now they go on to Describe the Backpropigation part. they explain it as it's known to us. However inturestingly they already mention a momentum factor that discounts the current Δw & can put more weight on the past value of w.

After all this they go on to show examples where a network with hidden layers & weight learned by their method can do what a simple input output network cannot.

They conclude by noting that this runs the risk of getting into local minima however inturestingly they state that through experimentation they found that this is only an issue when their "are just enough connections to perform the task" & that when more connections & dimentions are added they allow ways to escape the local minima.