INTERPRETING NEURAL NETWORKS

LECTURE 7
SECTION 2
JUNE 12TH

INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

UNIVERSITY of
RWANDA

WHAT DO NEURAL NETWORKS LEARN?

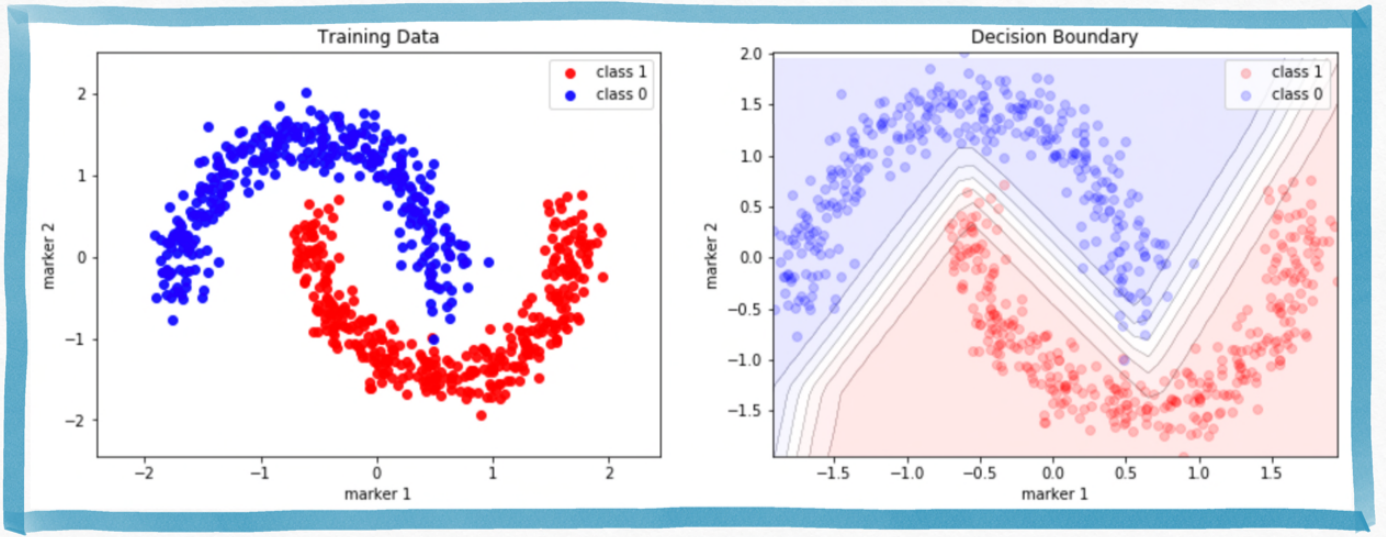
DECISION BOUNDARIES OF NEURAL NETWORKS:

Neural network models tend to outperform logistic regression. Why are they so effective?

An important way to understand neural network classifiers is to visualize their decision boundaries.

Complex Patterns in Data

Neural Networks Model Complex Boundaries

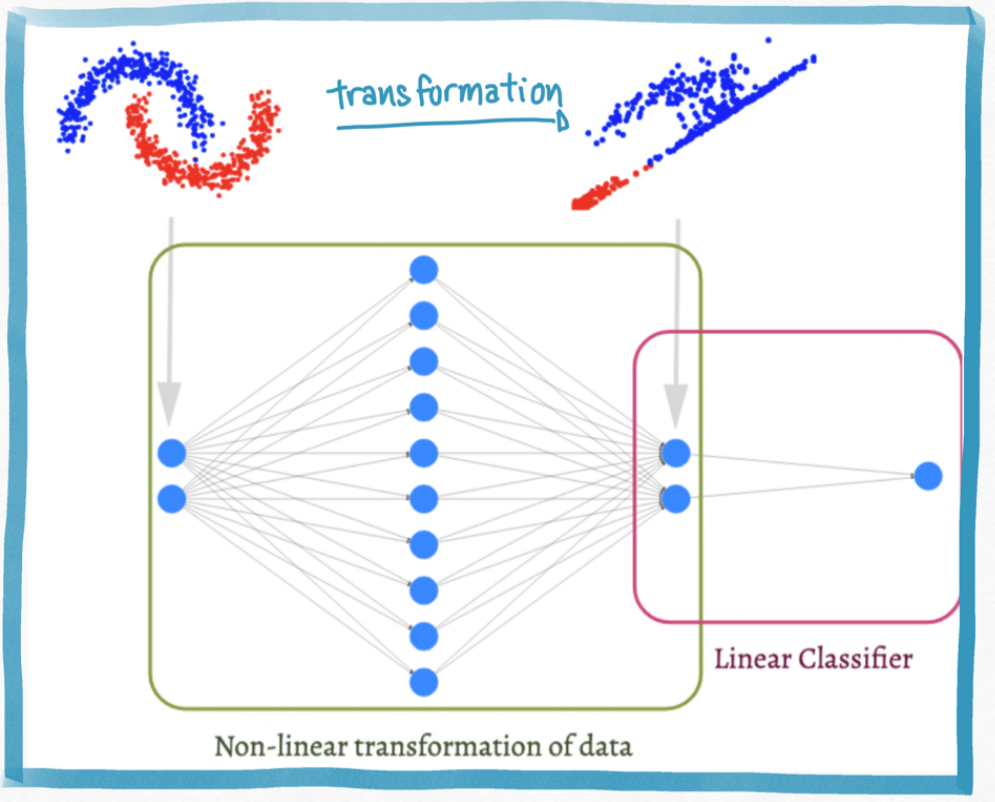


But visualization is only possible if the data is low dimensional!

HIDDEN LAYER REPRESENTATIONS:

Each hidden layer in the neural network contain different information about the data, these are called latent representations of the data.

At the input nodes, the data is not separable by a linear model.



By the last hidden layer, the data is transformed into something that is separable by a linear model.

The last layer in the neural network is just a linear classifier!

INTERPRETING NEURAL NETWORKS

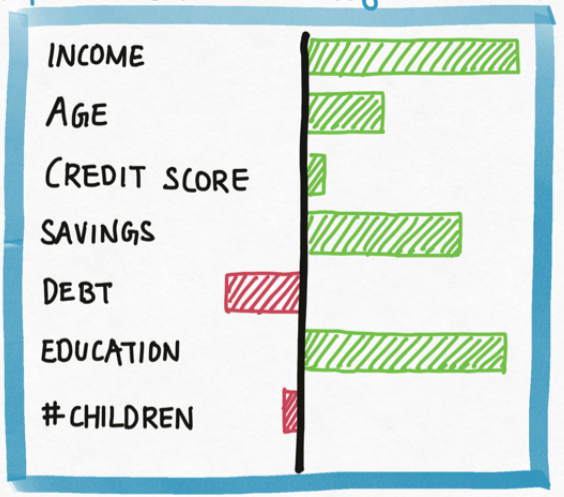
INPUT GRADIENTS:

When we interpret linear models, we look at all the parameters w in our model, there are as many parameters as there are covariates (plus the intercept).

Can we interpret the weights of neural networks?

We can interpret the gradient ∇_x at an input $x^{(n)}$, the input gradient tells us which input dimension has the biggest impact on the loss function.

Input Gradient of neg loss at $x^{(1)}$



Input Gradient of neg loss at $x^{(1)}$