

Philosophy & Ethics of Artificial Intelligence

Master in Philosophy of Science

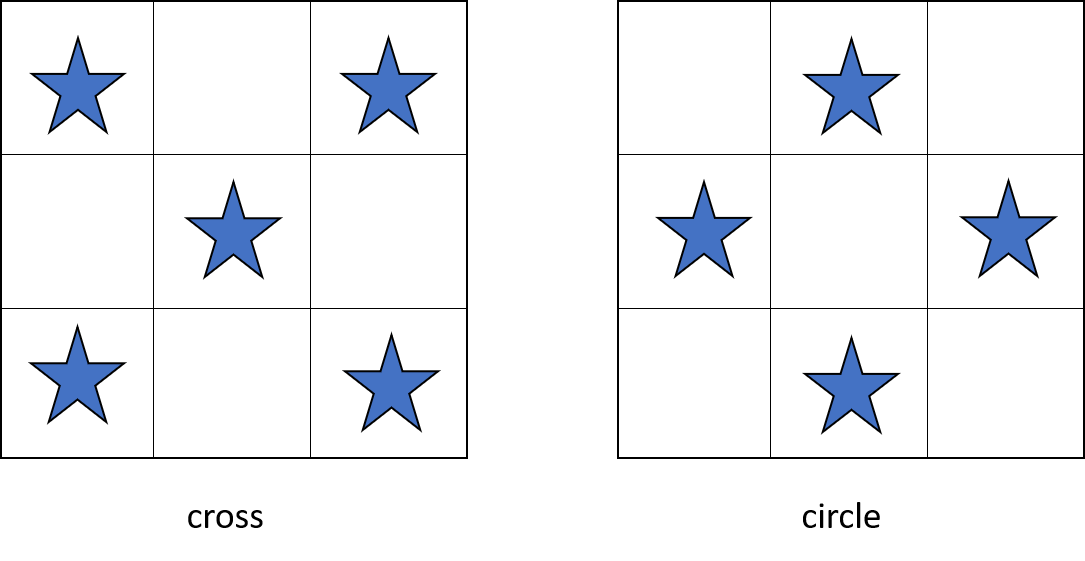
Philippe van Basshuysen

04 November 2019

Worksheet 3 – Epistemology of Machine Learning

1. **Recognising Simple Images Algorithmically**

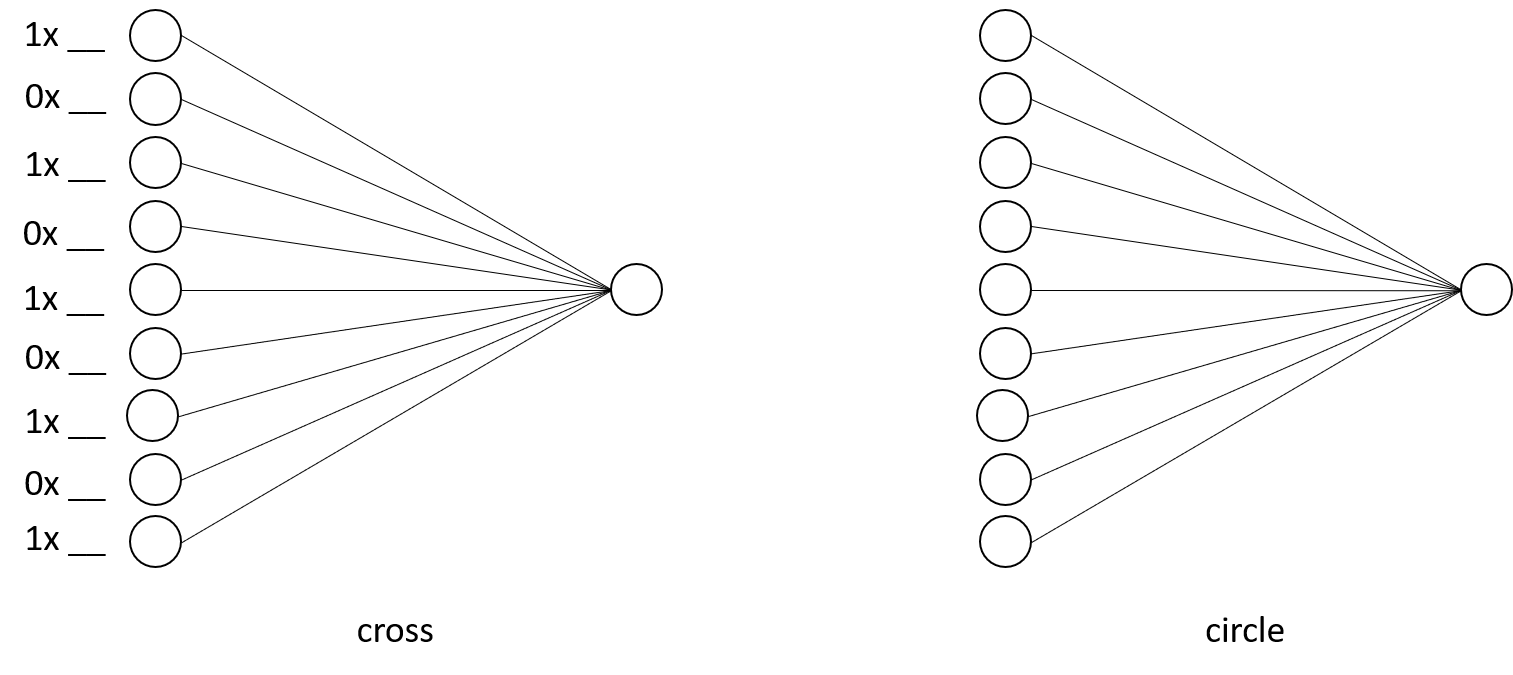
In the following images, suppose each square in the grid is a pixel. Stars denote coloured pixels and white squares denote white pixels. We wish to algorithmically recognise two shapes: a cross and a circle.



We choose numerical values for different pixels: white = 0, coloured = 1

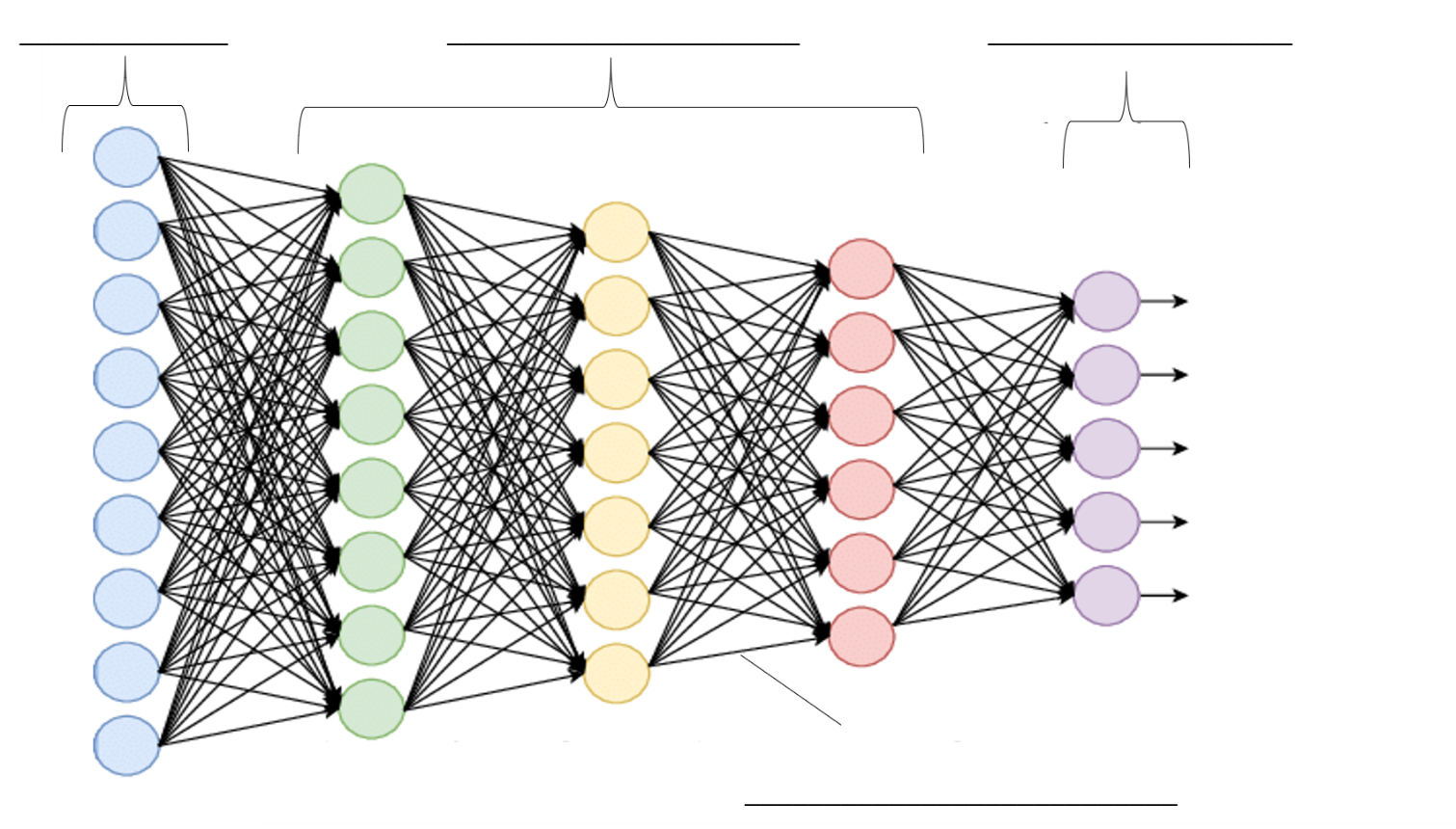
Express cross as “101010101”. Circle: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Determine the weights in the neural net below so that the following is the case: if the linear combination is <0 then the neuron activation is zero, which refers to cross. If the linear combination is >=0 then the neuron activation is one, which refers to circle.



1. **Deep Neural Networks (DNNs)**

What does a DNN consist in? Label the DNN below and describe in your own words what the different parts do.



How can a DNN be trained to recognise images (supervised learning)?

(Check <https://course.elementsofai.com/5/1> and <https://course.elementsofai.com/5/2> for a simple introduction to DNNs)

1. **Models**

What is link uncertainty and why does it occur?

Why are DNNs said to be “black-boxed”? Is this true? Are there ways to “open” the black box?

1. **Next Week: Risk and Potential of AI – The Possibility of Superintelligence**

Readings:

Stuart Russell (2019). *Human Compatible: AI and the Problem of Control*. Allen Lane. Chapter 3: HOW MIGHT AI PROGRESS IN THE FUTURE? and Chapter 5: OVERLY INTELLIGENT AI

Kevin Kelly (2017). The Myth of a Superhuman AI. <https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/>

Optional readings:

Nick Bostrom (2014). *Superintelligence*. Oxford: Oxford University Press. Chapter 4: THE KINETICS OF AN INTELLIGENCE EXPLOSION