2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 12 – 回推 DNN 模型輸入資訊

大綱

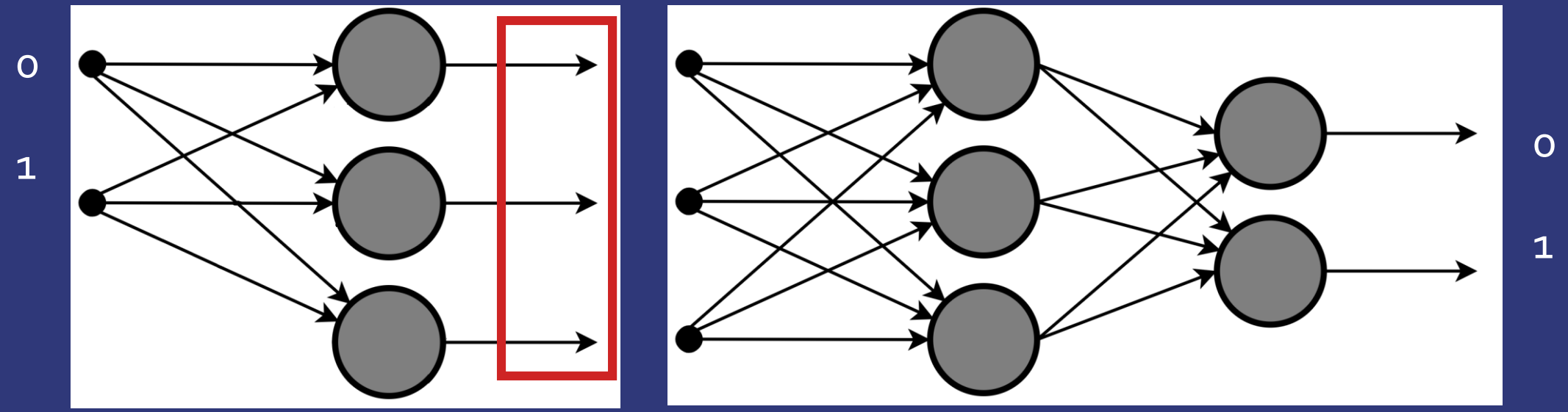
- 回推 DNN 模型輸入資訊
  - 前情再提要
  - 資料標準化、正規化
  - 程式實作
- 結論

前情再提要

- 如果攻擊者模型參數能夠讓 [0 1] 輸出結果為 [0 1]，就代表著把紅色框的結果灌入應用程式模型會得到 [0 1]
- 也就是說目前攻擊者模型對於 [0 1] 的輸出結果就是回推出的答案

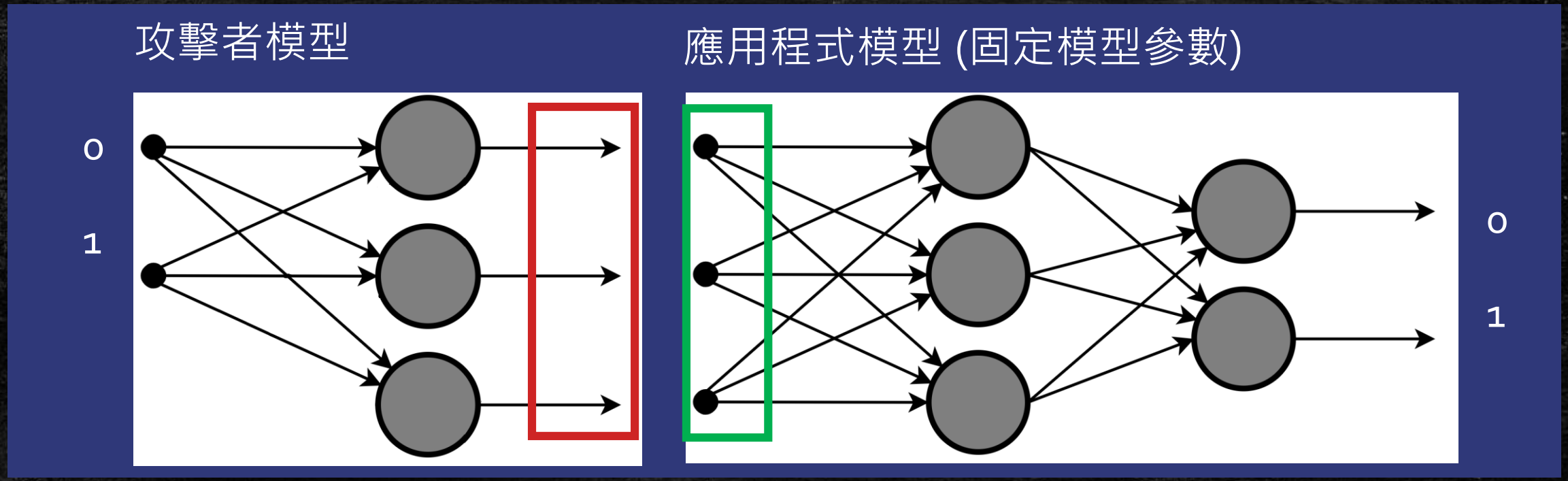
攻擊者模型

應用程式模型 (固定模型參數)



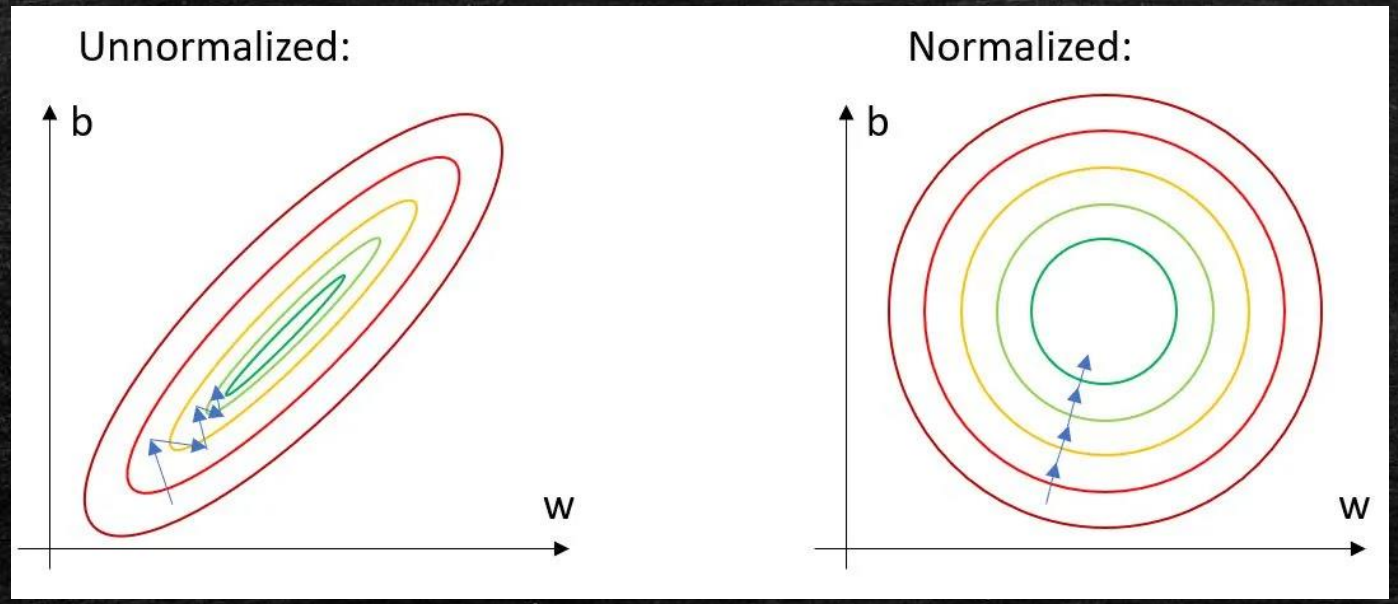
思考一下哪邊怪怪的?

- 當初綠色框框的輸入資料是怎麼來的?
- 28 × 28 的圖片資料，轉為 1 × 784 的向量，轉為浮點數在除上 255
- 所以只是輸出資料忘記乘上 255 還原而已



資料標準化、正規化

- 特徵標準化(normalization)是將特徵資料按比例縮放，讓資料落在某一特定的區間，可能是除上某個數值或是扣去平均值除上標準差



https://ithelp.ithome.com.tw/articles/10197357

程式實作

```
hack_data = attack_model.predict(final_target[0:2])
hack_img = np.reshape(hack_data[0],(28,28))

# https://ithelp.ithome.com.tw/articles/10197357?sc=rss.iron
hack_img = hack_img*255
```

```
hack_img = hack_img.astype(np.uint8)
im = Image.fromarray(hack_img)
im.save("hack.bmp")

test_data = np.asarray(hack_img)
plt.imshow(test_data,cmap='gray')
```

結論

- 細節藏在細節裡，想不到一個常見的資料標準化影響了整個攻擊流程

- 回到攻擊者的角度，所以知道模型結構跟參數還是不夠的，還需要知道當初開發者用了哪種資料標準化的運算才能夠回推出真實的輸入資料