機器學習 Final Project toy model

劉昱賢

Contents

1 toy model 設置

1.1 玩具模型簡化想法

想法：

輸入聲波以及標籤，將聲波均分成許多點並丟進 *nerual network* 中，最後求出標籤爲 *0, 1*，並驗證結果。

1.2 資料設定

爲了簡化問題，將資料設定爲 $sinx$ 的波，週期爲 2π 。每個波的能量、週期均相同。

每個樣本的正弦波：$s_i(x) = \sin(x + \phi_i), \quad i = 1, 2, \ldots, 5000, \quad x \in [0, 2\pi), \quad \phi_i \sim \textit{Uniform}(0, 2\pi)$

檢視最高點的位置，若最高點在 $(0, \pi)$ 之間，則將資料標示爲 1。

若最高點在 $(\pi, 2\pi)$ 之間，則將資料標示爲 0。

除此之外，會對不同情況進行討論，測試不加 *noise* 是否可行，接著再加上 *noise* 模擬不同狀況。共生成 *5000* 筆資料。

1.3 模型介紹

先將資料切分爲三個部分，*Training Set, Validation Set, Test Set*，分別占比例 *80%, 10%, 10%* 。

每個聲波會先拆分成 *1000* 個點，每個點間隔均相同。接著將聲波的點放入 *Neural Network* 中進行訓練，最後得出分類狀況。

Input: $x \in \mathbb{R}^{1000}$ *, Label:* $y \in \{0, 1\}$ *, Output: 0 or 1* (*Binary Classification*)

Model: $h_\theta(x) = \sigma\Big(W_2 \cdot \text{ReLU}(W_1 x + b_1) + b_2\Big)$ *,* 訓練次數 *3000* 次

$x \in \mathbb{R}^{1000}$ *,* $W_1 \in \mathbb{R}^{32 \times 1000}, \quad b_1 \in \mathbb{R}^{32}$ *,* $W_2 \in \mathbb{R}^{1 \times 32}, \quad b_2 \in \mathbb{R}$ *,* σ *is Sigmoid function*

Loss function: BCE loss

評估方法：*Confusion Martix + accuracy*

2 實作與結果

2.1 不加 noise

生成並繪製了 *5000* 個波形，並將所有結果繪製。紅線代表標籤 *1*，藍線代表標籤 *0*。

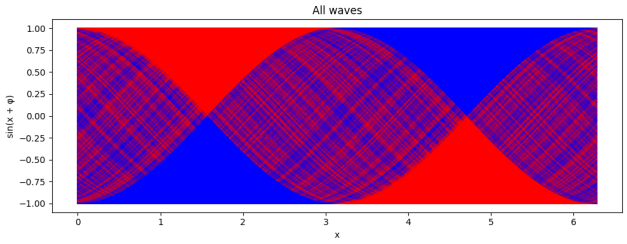

Figure 1: All wave with no noise

在生成完 *5000* 個波形後，我們使用神經網路對數據進行訓練，觀察在無噪音的理想情況下的模型表現。訓練 *3000* 次迭代後，結果如下：

- *Train loss = 0.0007*

- *valid loss=0.0024*

對應的混淆矩陣如下：

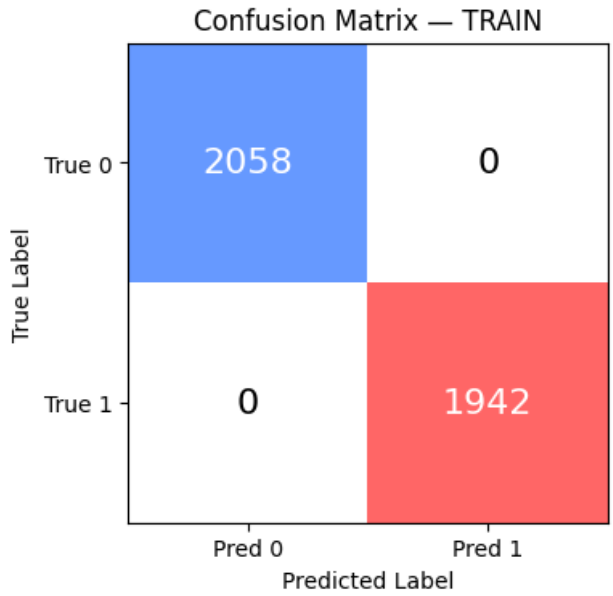

Figure 2: Confusion Matrix in train data with no noise

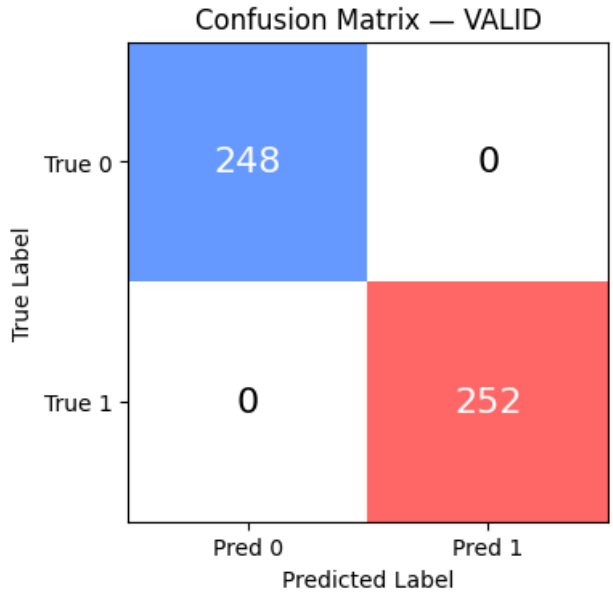

Figure 3: Confusion Matrix in valid data with no noise

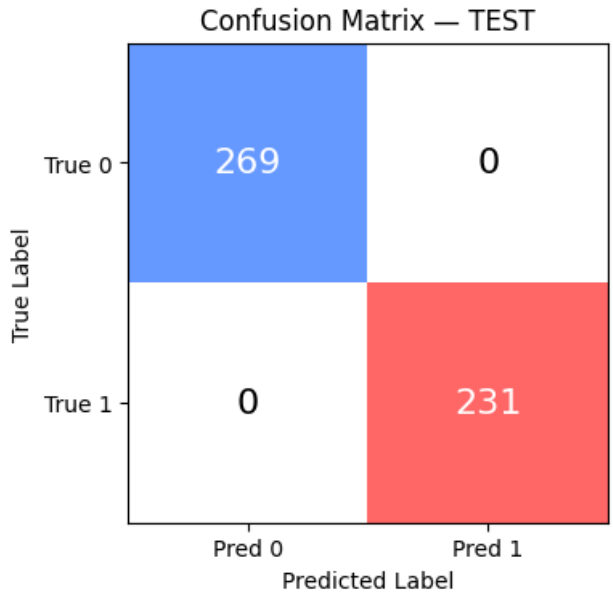

Figure 4: Confusion Matrix in test data with no noise

三個資料集的準確率如下：

$$\textit{Train Accuracy} = 1.000, \quad \textit{Valid Accuracy} = 1.000, \quad \textit{Test Accuracy} = 1.000$$

訓練結果顯示，無噪音情況下模型能夠完美分類所有波形，三個資料集的準確率皆達 *100%*。

2.2 noise=0.3

我們生成並繪製了 *5000* 個波形，每個點上都加入了 *noise*。

紅線代表標籤爲 *1* 的波，藍線代表標籤爲 *0* 的波。

每個波的 *1000* 個點都加入了隨機噪音：$s_i(x_j) = s_i(x_j) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, 0.3^2), \quad i = 1, \ldots, 5000, \quad j = 1, \ldots, 1000$

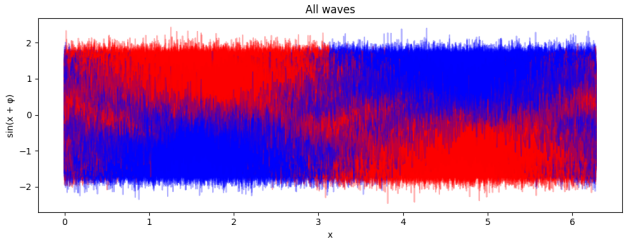

Figure 5: All wave with noise=0.3

單獨觀察第一個波型，如下

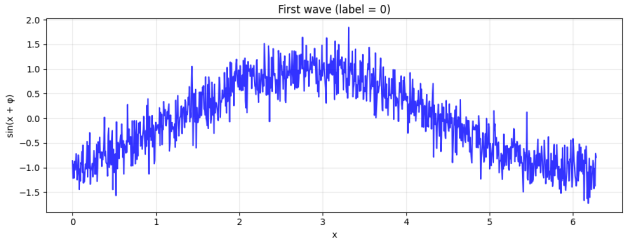

Figure 6: First wave with noise=0.3

在生成完 *5000* 個波形後，我們使用神經網路對數據進行訓練，觀察在加入噪音的情況下模型的表現。

訓練 *3000* 次迭代後，結果如下：

- *Train loss = 0.0001*

- *Valid loss = 0.6151*

對應的混淆矩陣如下：

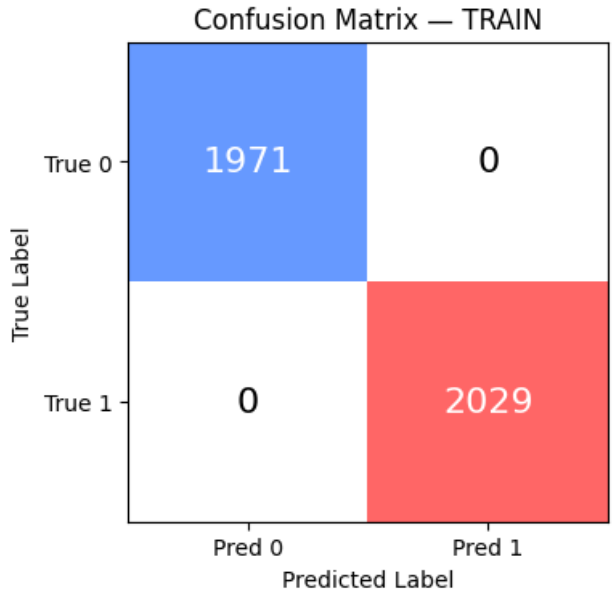

Figure 7: Confusion Matrix in train data with noise=0.3

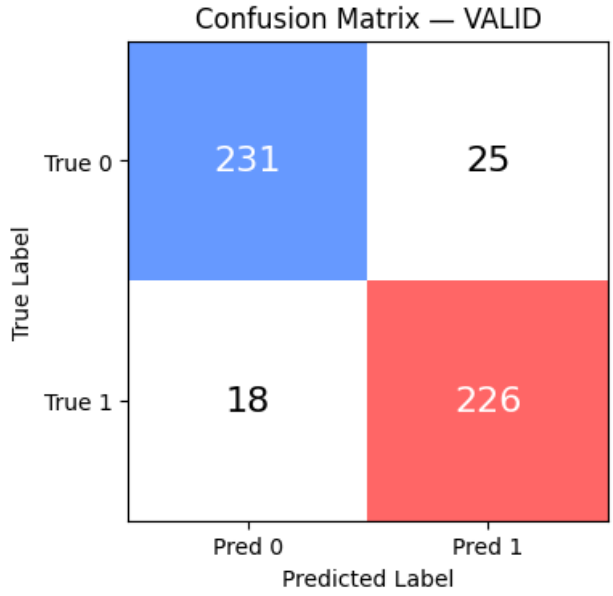

Figure 8: Confusion Matrix in valid data with noise=0.3

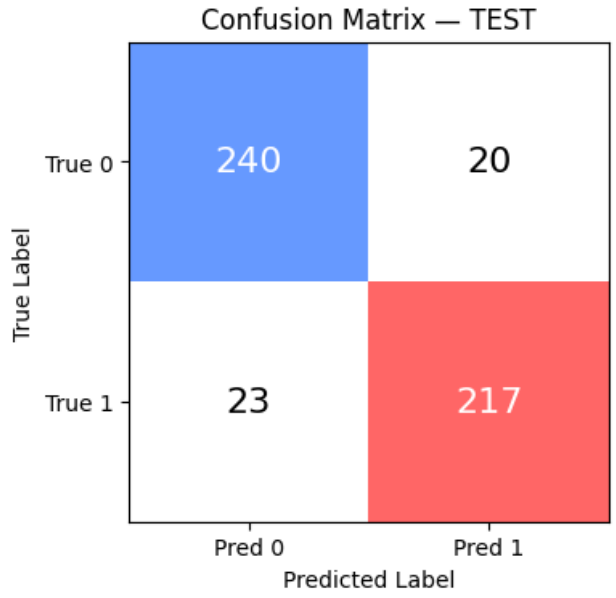

Figure 9: Confusion Matrix in test data with noise=0.3

三個資料集的準確率如下：

$$\textit{Train Accuracy} = 1.000, \quad \textit{Valid Accuracy} = 0.914, \quad \textit{Test Accuracy} = 0.914$$

訓練結果顯示，在加入噪音後，模型在訓練集仍可達到完美分類，但在驗證集與測試集的準確率下降到約 *91.4%*，顯示噪音對模型造成了一定影響。

2.3 noise=0.6

由於在噪音 $\textit{noise\_std} = 0.3$ 時訓練結果仍然不錯，我們進一步測試 $\textit{noise\_std} = 0.6$ 的情況。

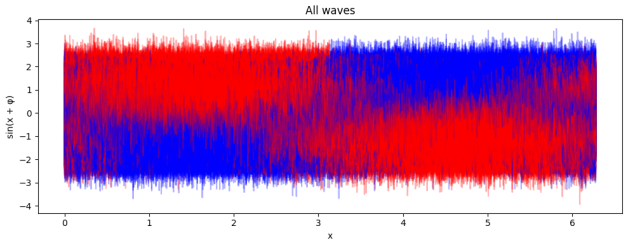

Figure 10: All wave with noise=0.6

單獨觀察第一個波型，如下

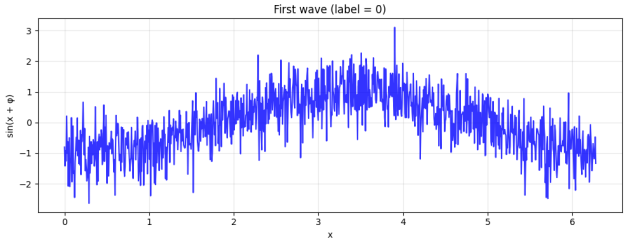

Figure 11: First wave with noise=0.6

在生成完 *5000* 個波形後，我們使用神經網路對數據進行訓練，觀察在加入較大噪音的情況下模型的表現。

訓練 *3000* 次迭代後，結果如下：

- *Train Loss = 0.0001*

- *Valid Loss = 1.8199*

對應的混淆矩陣如下：

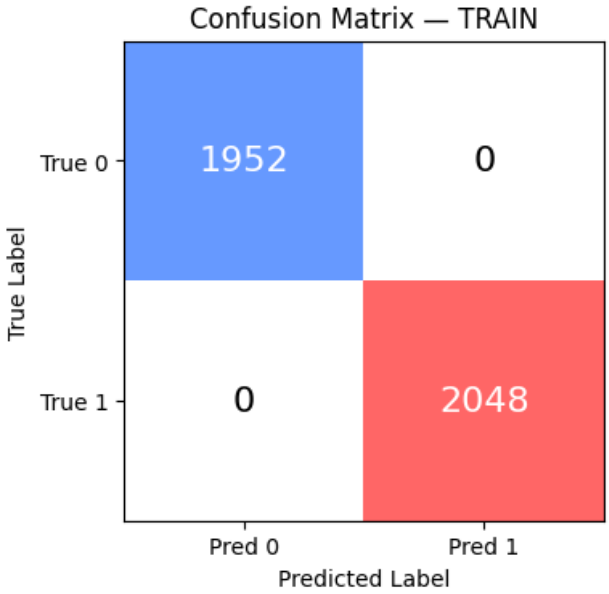

Figure 12: Confusion Matrix in train data with noise=0.6

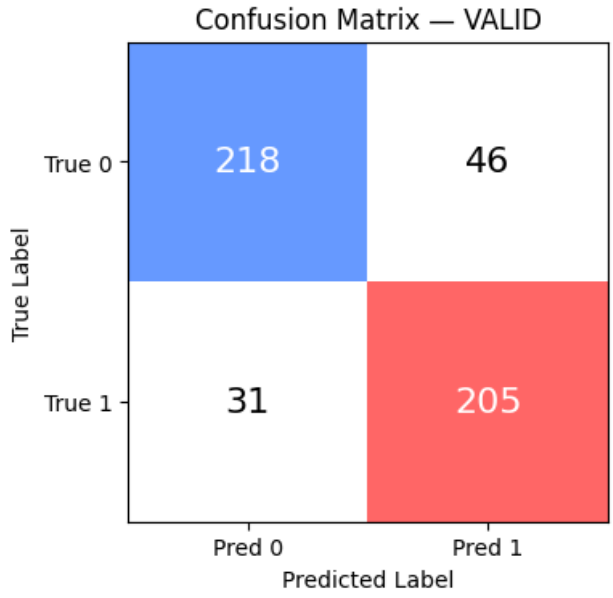

Figure 13: Confusion Matrix in valid data with noise=0.6

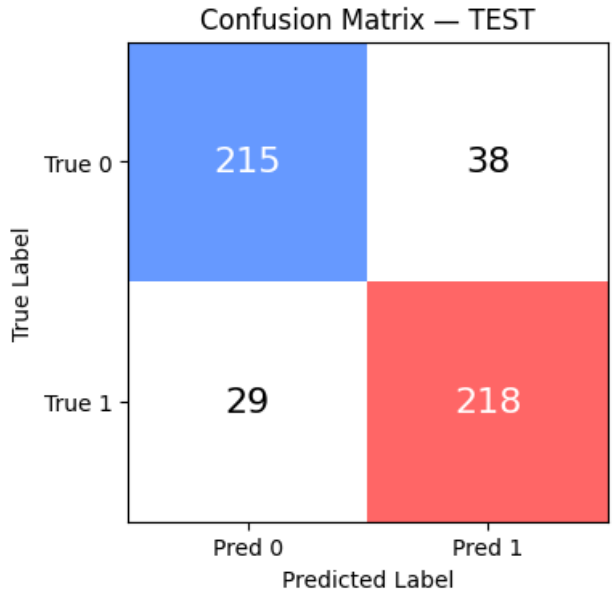

Figure 14: Confusion Matrix in test data with noise=0.6

三個資料集的準確率如下：

$$\textit{Train Accuracy} = 1.000, \quad \textit{Valid Accuracy} = 0.846, \quad \textit{Test Accuracy} = 0.866$$

訓練結果顯示，在較大噪音的情況下，模型雖然在訓練集仍可達到完美分類，但在驗證集與測試集的準確率下降，顯示噪音對模型產生了較大影響，但結果依然還不錯。

3 結果討論

3.1 噪音對模型表現的影響

當 $noise = 0$ 時，模型完全學會了數據特徵，三個資料集的準確率皆爲 *100%*。

當 $noise = 0.3$ 時，訓練集仍保持完美分類，但驗證集和測試集下降到約 *91.4%*，說明噪音增加了資料的變異性，使模型在未見過的資料上分類能力略微下降。

當 $noise = 0.6$ 時，驗證集與測試集下降到約 *85—87%*，顯示噪音過大會顯著影響模型能力，雖然訓練集仍可完美分類（可能存在過擬合現象）。

3.2 訓練 Loss 與驗證 Loss 的差異

當 $noise = 0$ 時，*Train Loss* 與 *Valid Loss* 均接近 *0*，顯示過擬合問題不存在。

當 $noise > 0$ 時，*Train Loss* 仍很低，但 *Valid Loss* 明顯上升，說明模型對訓練資料中的噪音過度擬合，而驗證資料時出現誤差。

3.3 判斷錯誤的樣本特徵

下圖顯示在 $noise = 0.3$ 與 $noise = 0.6$ 時被誤判的波形樣本。

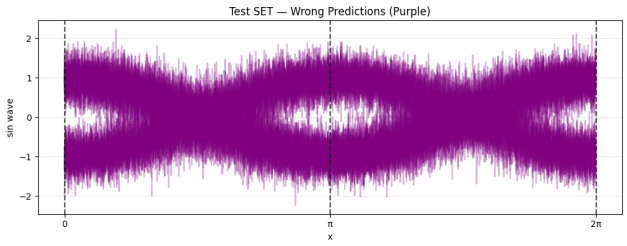

Figure 15: Wrongly predicted wave with noise=0.3

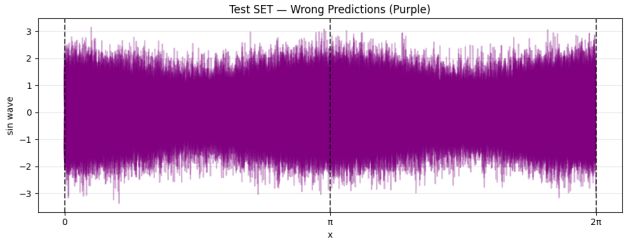

Figure 16: Wrongly predicted wave with noise=0.6

觀察發現，在有噪音的情況下，容易判斷錯誤的資料及其最大值會集中在 $0, \pi, 2\pi$ 附近，這些位置對分類來說較爲模糊，容易受到噪音干擾。

4 結論

1. 模型對原始波形的學習效果非常好，能夠完美分類無噪音資料。
2. 隨著噪音增加，模型在訓練集上的表現仍然良好，但驗證集與測試集的準確率逐漸下降，顯示噪音對模型具有明顯影響。
3. 當噪音增加時，訓練 *Loss* 仍保持極低，但驗證 *Loss* 明顯上升，可能是模型對訓練資料的噪音過擬合所致。
4. 過大的噪音會降低驗證集與測試集的分類效果，但模型仍能保留一定辨識能力。
5. 若希望模型在更大噪音下穩定，可考慮增加資料量或使用更深層網路等方法。