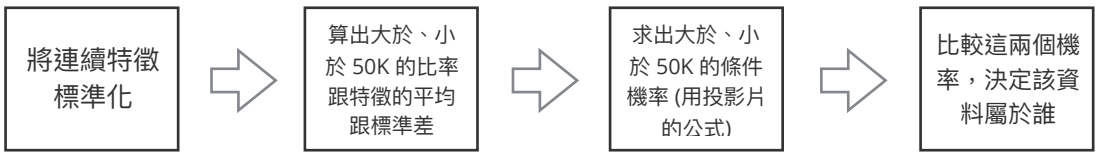
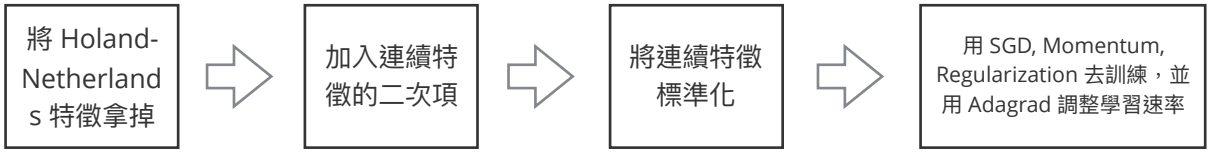
1. 請說明你實作的 generative model，其訓練方式和準確率為何？
答：



- 一開始我沒有將連續特徵正規化，結果在訓練資料上準確率只有 76%。
- 因為是按照投影片的公式去實作，因此是以高斯分布去描述特徵跟結果的關係。

	訓練資料	測試資料 (Kaggle public)
準確率	0.81278	0.83391

2. 請說明你實作的 discriminative model，其訓練方式和準確率為何？
答：



- 會將上面提到的特徵拿掉是因為，該特徵只有一筆資料有值，這導致該特徵對應到的 weight 特別大 (因為 Adagrad 的計算方法)
- 我有把 20% 的訓練資料當成驗證資料，用它們來當作判斷模型好壞的準則

	訓練資料	驗證資料	測試資料 (Kaggle public)
準確率	0.8553	0.8570	0.85811

3. 請實作輸入特徵標準化 (feature normalization)，並討論其對於你的模型準確率的影響。
答：

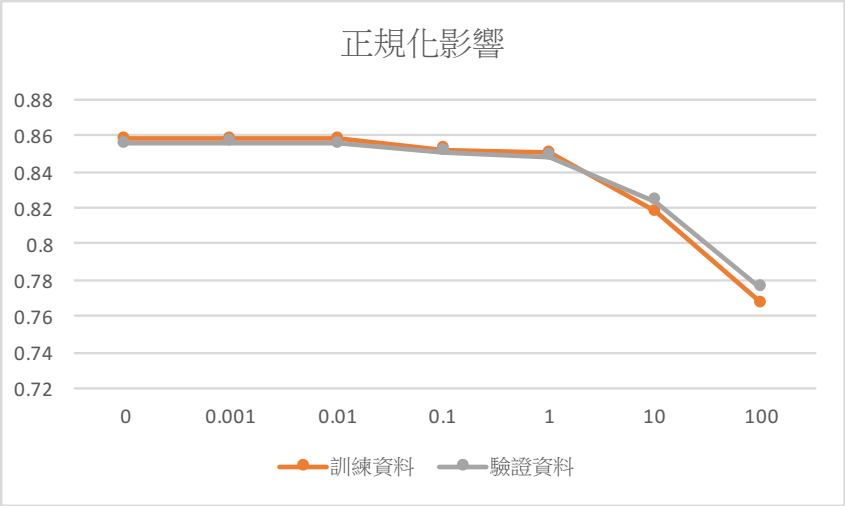
- 一開始實作 generative model 時，我並沒有做輸入特徵標準化，結果準確率在訓練資料上只有 76%。在加入輸入特徵後，準確率可以提升到 81%。
- 在實作 discriminative model 時，我甚至沒辦法在沒做特徵標準化時訓練，因為幾個連續資料的尺度都不一樣，很容易就出現 overflow。我一開始是對所有資料做標準化，雖然一樣能訓練，但結果一直無法超過 strong baseline，因為我把原本只有 1 跟 0 的資料都拉近了，失去原本資料包含的意義。
- 輸入特徵標準化之所以能提升準確率是因為，這個動作能把原本不同尺度的特徵，在不影響其代表意義之下，將它們拉到同一個尺度，使在調整參數時較容易到達 loss 的低點。

4. 請實作 logistic regression 的正規化 (regularization)，並討論其對於你的模型準確率的影響。

答：

我設計一個實驗，藉由調整正規化的參數，去觀察其對準確率的影響，結果如下表：

	準確率	
正規化參數 λ	訓練資料	驗證資料
0 (沒有正規化)	0.8577	0.8555
0.001	0.8578	**0.8557**
0.01	0.8576	0.8554
0.1	0.8522	0.8505
1	0.8501	0.8480
10	0.8176	0.8236
100	0.7670	0.7755



參數設為 0.001 時，在驗證資料上有最好的準確率，但在訓練資料上也有最好的準確率。這跟預期的結果不太相符，因為正規化應該是會讓訓練上的結果變差，驗證上的變好。我猜測是因為正規化在這次的作業中影響並不大，原因是很多資料不是連續的，使得正規化將曲線平滑化的效果表現不出來。

5. 請討論你認為哪個 attribute 對結果影響最大？

答：

我從最好的模型的 weight 去判斷，將每個特徵對應到的 weight 按照大小排序，結果如下：

age > capital_gain > Doctorate > Prof-school > Wife

前兩名的差距並不大 (2.22, 2.12)，而且屬於連續的 attribute，從此可以推論出連續的 attribute 在分類上較有幫助，因為它們更能把兩個種類區分開來。

結論：age 跟 capital_gain 對結果影響最大。