學號:R07941023 系級: 光電碩二 姓名:呂彥穎

1. (0.5%) 請比較你實作的 generative model、logistic regression 的準確率,何者較佳? Sol:

Kaggle (logistic regression):0.85429/0.84817

Kaggle (generative model):0.84324/0.84375

logistic regression 的效果比較好,其原因 generative model 是一個 data 彼此獨立的機率模型,在 data 少的時候具有一定程度的腦補想像力而準確度上升,但當 data 數量夠大的時候,logistic regression 減少 bias 帶來得誤差後,其準確度會大於 generative model。

2. (0.5%) 請實作特徵標準化(feature normalization)並討論其對於你的模型準確率的影響標準化的目的是讓多維的 feature 是個圓形,可以快一點到達 local minimum,如果時間是無限的,準確率不應該受到影響。

固定所有參數,只考慮 epochs:100 (注意:兩個都還沒收斂,僅只比較而已)

Original feature

Kaggle (logistic regression):0.23525/0.23719 (全都還是 1,收斂超慢)

Feature normalization

Kaggle (logistic regression):0.8371/0.8349

3. (1%) 請說明你實作的 best model, 其訓練方式和準確率為何? Sol:

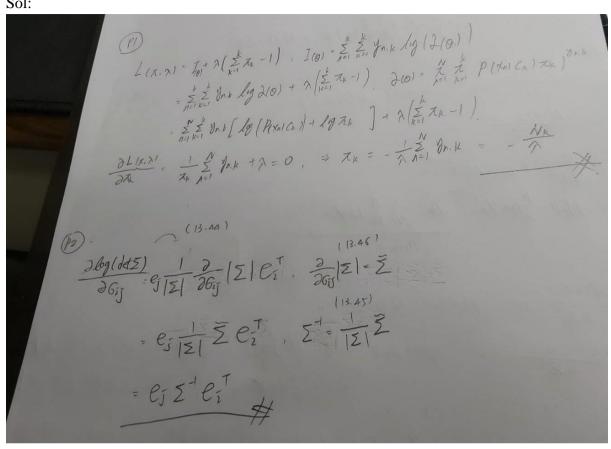
Kaggle (best model):0.86953/0.85468

這裡使用的是 gradientboosting。

訓練步驟:

- 1.調個我破電腦不會跑太慢的 learning rate,這裡先設個 0.1
- 2. 寫個迴圈去跑最低 loss 需要的"樹"數量, 跑完後選擇 400 顆。
- 3.重複步驟 2, 去寫迴圈單變數測試 depth/split/left/subsample
- 4.經過一番努力後,得到這些參數 depth = 8/split=50 /left=7/subsample=0.8
- 5.可靠資料說,這時候就可以降低 larning rate10 倍,然後讓樹增加 10 倍,結果壞掉了,所以回到上一步就直接跑分。

4. (3%) Refer to math problem Sol:



$$P(x) = \frac{1}{2} \int_{-\infty}^{\infty} \frac{dx}{dx} \int_{-\infty}^{\infty} \frac{$$