ML 2019 SPRING HW2

B06902074 資工二 柯宏穎

1. 請比較你實作的generative model、logistic regression 的準確率，何者較佳？

在完全沒有加入其他優化的情況下，generative model有較好的準確率，但單純靠distribution下去training，跟logistic regression相比，變化性相對較小，我加入了二次項，標準化等，其大致的分布仍不會有太大的改變，準確率並沒有明顯的成長。(表格內的精確度為public/private)

	Origin	Normalization&non-linear
Generative model	0.83108/0.82397	0.84447/0.84535
Logistic regression	0.79004/0.78503	0.85847/0.85530

2. 請說明你實作的best model，其訓練方式和準確率為何？

我並無使用其他訓練方式來做taining，也因為如此我並無超過private的strong。我只透過調整logistic regression來提高準確度。我加入了normalization，regualarization，non-linear regression加到四次項。訓練時我會隨機抓取一個人來調整我的weight，這會使得準確度有些浮動，但在大量iteration下是差不多的。這次的訓練資料並不算大量，較不會有overfiting的問題，public和private的精確度皆差不多。最好的一次準確度為0.85530/0.85847(piblic/private)。

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

在沒加入normalization的model裡，精確度是較差的。我們所抽取的資料中，每種的變化幅度不同，且有些數值偏大，在算sigmoid時常會爆掉，特判掉後通常會是一大堆的0。沒做標準化的精確度 0.79004/0.78503， 有做的 0.85147/0.85345。

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。

加入正規化後，防止其衝入極值點造成overfitting。不過這次數值有些偏大，λ不能調太大，我的λ抓在10^{-6}左右，避免regualarization的影響超過logistic regression的計算。沒做正規化的精確度 0.79004/0.78503， 有做的 0.84115/0.83269。

5. 請討論你認為哪個attribute 對結果影響最大？

我用加入標準化且沒有使用non-linear regression那次的來看，我抓取weight權重值最大(positive) 與最小(negtive)的來判斷。發現為capital_gain與white。資本收入越高的人，收入超過一定水準的機率較大，算滿合理的。white雖然為最小，但與black相比其實差不多小(-1.682, -1.458)，表示不同人種對最後的影響並無絕對的關係。主要還是以capital_gain的影響最大。