Homework 2 Report - Income Prediction

學號：b03902130 系級：資工四 姓名：楊書文

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

MODEL	PUBLIC	PRIVATE
Logistic (only scaling)	0.82125	0.81353
Logistic (feature engineering)	0.86007	0.85456
Generative (only scaling)	0.84643	0.84289
Generative (feature engineering)	0.84766	0.84166

從上表可以看出 logistic model 在 feature preprocessing 之前比 generative 差了許多，但 preprocessing 之後，又比 generative 好了許多。我想是因為 logistic 是直接照著 training data 的樣子去學，當 data 間的關係不明確時，就難以學到有用的資訊。因此 train 的過程中，一找到對的 preprocessing 方法，accuracy 就能明顯的 train 上去。而 generative model 直接對資料的分布做假設，就可以避免 data 不乾淨時學到無用的資訊，然而這假設同時也限制了 generative 的 accuracy 上限，因此儘管用了處理過的 data 仍然沒什麼進步。

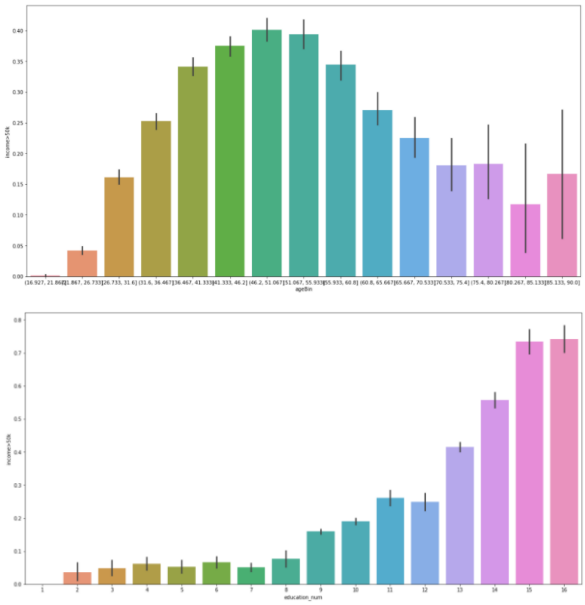
(1%) 請說明你實作的 best model，其訓練方式和準確率為何？

我的 best model 是 logistic model with adagrad，不過有特別做 feature 處理，加了一些 new feature。分別有：

capital_gain 平方：因為做圖發現 capital_gain 在 10000 之前等區間分隔後，每個區間中的人會致富的條件機率越往大數值(右邊)的區間就會越大，且成指數關係，因此加入平方項。

hours_per_week > 39：同樣的，在區間的條件機率上，約從 40 小時以上的機率較高，以下則低很多。因此加入是否大於 39 的 one-hot (valid 上 39 的分數比 40 好。)

1 / abs(age – 50)：因為發現年齡與致富的機率關係約成右圖，最高的機率在中間，因此試圖加入一個能反映機率的新 feature。圖中最高點為 50。



train.csv 中的 education_num 平方：因為助教抽的 feature 中 ，教育時間被當作 one-hot 處理，然而該數值其實和致富機率很相關且約成指數(如右下圖)，因此 選擇加入其 raw data 的平方項。

capital_gain 的區間 one-hot：因為 gain 雖然 10000 之前有指數關係，但 10000

之後資料很少且非指數，因此加入 gain 的 one-hot 試圖讓 model 學到不同的區間怎麼分別處理。

(1%) 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。(有關 normalization 請參考：https://goo.gl/XBM3aE)

MODEL	PUBLIC	PRIVATE
with normalization	0.79975	0.79044
without normalization	0.81068	0.80444

從表上可以看到，沒有 normalization 的影響是相當大的。我想這是因為 logistic model 畢竟是 non-linear model，不像作業一的 linear convex 那麼好走，因此如果沒有讓 feature 間的數量級一致，error surface 就會較為扭曲，加上 adagrad 也不是特別強的動態步伐 optimizer，因此就較不容易走到最低點。

(1%) 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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

MODEL	PUBLIC	PRIVATE
Regularization term = 0	0.81068	0.80444
Regularization term = 1	0.83034	0.82889

可以看到加了 regularization 是有明顯變好的，這和作業一相當不同，作業一中加了 regularization term 通常會變差。我想可能也是因為 non-linear error surface 的關係，當 surface 為單純 convex，該走的方向很明確時，加了 regularization 是限制走的速度，較難走到最低點。而在 non-linear error surface 時，走的方向就變得很重要，一旦走偏可能會完全走不到最低點，因此當我一加了 regularization=1，走的速度看起來比沒加快很多，代表應該走得很順，沒有卡到什麼 local minimuma。

(1%) 請討論你認為哪個 attribute 對結果影響最大？

MODEL	PUBLIC	PRIVATE
drop sex	0.81953	0.81157
drop capital gain	0.80601	0.80248
drop hours per week	0.81633	0.81525
drop age	0.81830	0.81058
drop relationship	0.82358	0.81697
add (capital gain – capital loss)	0.82457	0.81673

用 raw data 做圖看各 feature 與致富的關係圖，列出幾個比較相關的，並比較把他們各自 drop 掉後對 model 的影響。根據上表可以發現，把 capital gain drop 掉後對 performance 的影響最大。同時一旦加了新 feature: (capital – loss)，performance 也有提升。因此我認為 capital_gain 對效果的影響最大。