學號：B03902042 系級： 資工三 姓名：宋子維

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

答：這次的作業是binary classification，所以訓練方式即為：

I. 假設兩類的訓練資料分別由Gaussian Distribution產生出來，且此分布擁有maximum likelihood。

II. 分別算出兩個類別 $\{\leq 50K,\ > 50K\}$ 的mean和covariance。

III. 對於測試資料 x ，用剛剛求出的 Gaussian Distribution 算出 likelihood，因此能求出 $P(\leq 50K \mid x)$ 和 $P(> 50K \mid x)$ ，看哪個機率大就將 x 分到哪類。

我使用的是助教預先抽出的 feature，並對 feature 做 z-normalization，在public set上的準確率為0.84115。

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

答：實作的 model 為 logistic regression model，訓練方式為 full batch gradient descent，詳細如下：

I. 初始化 weight w_i 和 bias b。

II. *for epoch in* [1, *max_epochs*]

A. 對於訓練資料 $(x,\ y)$ ，計算預測值 $\hat{y} = sigmoid(wx + b)$。

B. 計算 loss function cross entropy: $y\ \log\hat{y} + (1-y)\log(1-\hat{y})$ 的gradient $g_{w_i} = -(y-\hat{y})\,x_i$ ， $g_b = -(y-\hat{y})$。

C. 更新參數： $w_i \leftarrow w_i - \eta g_{w_i}$ ， $b \leftarrow b - \eta g_{w_i}$。

和第一小題相同，一樣使用助教預先抽出的feature，且對feature做z-normalization，訓練次數為2000次，初始學習率為0.05，並使用adagrad，在public set上的準確率為0.85369。

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：**原本完整的 data 有32562筆，我將最後6000筆當作 validation set，因此實際用來訓練的資料有26562筆**，而 feature 一樣是直接使用助教預先抽好的，訓練次數為10000次，初始學習率為0.005，並且使用adagrad。

Normalization?	Training (accuracy/loss)	Valid (accuracy/loss)
Yes	0.83935/0.37693	0.84333/0.37430
No	0.79843/0.46212	0.80450/0.44969

可以看到在相同的訓練次數和初始學習率之下，feature normalization 能大幅提升準確度和效能。另外，若沒有使用 feature normalization ，即使用了adagrad，初始學習率也必須好好選擇，否則容易在訓練過程中發散。

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：下表中的"origin"為助教預先抽取好的 feature，"square"為age，fnlwgt，capital gain，capital loss，和 hours per week，這五個欄位的平方項。

Regularization constant	Feature	Training (accuracy/loss)	Valid (accuracy/loss)
0.0	origin	0.85309/0.31678	0.85167/0.31295
0.001	origin	0.85354/0.32741	0.85250/0.32378
0.0001	origin	0.85317/0.31800	0.85167/0.31417
0.0	origin+square	0.85686/0.30798	0.85600/0.30761
0.001	origin+square	0.85584/0.32655	0.85633/0.32474
0.0001	origin+square	0.85693/0.31079	0.85600/0.31017

可以看到加了regularization後，對模型準確率的影響並不大，也許是feature的選擇，使得模型並沒有太過複雜，因而不會與訓練資料過度擬合，產生overfitting。

5. 請討論你認為哪個attribute對結果影響最大？

答：**我認為capital gain的影響最大**，我有試過將各項 continuous 的 attribute 進行轉換，如 log() ，*n* 次方等，而從實驗數據來看，加上 log(*capital gain*) 後，不只模型收斂速度快很多，在 public set 上也能取得較好的成績。若單就一次項模型進行比較，試著個別將各項attribute拔掉後，並觀察其結果，可以發現移除capital gain後的結果最差。

Removed attribute	Training (accuracy/loss)	Valid (accuracy/loss)
None	**0.85309/0.31678**	**0.85167/0.31295**
age	0.85053/0.32092	0.85433/0.31694
fnlwgt	0.85234/0.31743	0.85250/0.31374
sex	0.85313/0.31927	0.85167/0.31432
capital gain	**0.83807/0.34487**	**0.83783/0.34542**
capital loss	0.85125/0.32179	0.85150/0.31882
hours per week	0.85046/0.32241	0.85300/0.31894
workclass	0.85076/0.31887	0.85183/0.31633
education	0.84233/0.33314	0.84400/0.33075
marital status	0.85339/0.31871	0.85200/0.31590
occupation	0.84718/0.32614	0.84917/0.32187
relationship	0.85336/0.31851	0.85050/0.31441
race	0.85328/0.31738	0.85183/0.31375
native country	0.85181/0.31885	0.85050/0.31267

（註：hw2_best.sh中，我實作了Gradient Boosting，以CART作為weak learner。）