學號：B04902099 系級： 資工三 姓名：黃嵩仁

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

	Training set accuracy	Testing set accuracy
Generative model	0.842203	0.843744
Logistic model	0.853567	0.852957

在我所實作的 generative model、logistic regression 兩個 model 中，logistic model 的準確率較佳，可能的原因是，所提供的資料分布並沒有很好的與 generative model 假設的機率分布一致。
註：
(1)上方成績之 generative model 所用之 covariance matrix 為 Y_train == 0 與 Y_train == 1 所分別算出之 covariance matrix 加權平均所得
(2) 上方成績之 logistic model 的實作方法為 adagrad，lr = 0.05，epoch = 3000
(3)為了公正性，上方得到的數據，兩種 model 都採用助教所提供的 feature，並經過 z-normalization(上傳至 github 的 logistic model 有再增減 feature，並調整其他參數)。

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：實作的 best model 使用 xgboost 套件，XGBoost(eXtreme Gradient Boosting)是 Gradient Boosting 算法的一個優化的版本，背後使用了 gradiant boosted trees(boosting 使用了 ensemble 的概念)。
在嘗試不同的參數組合後，最後決定參數為 max_depth=3， n_estimators=1500，learning_rate=0.05

	Training set accuracy	Testing set accuracy
準確率	0.886305	0.877218

註：使用助教預先抽取出的 feature 檔作為 input，刪減部分 feature、再將部分連續的 feature 取次方後，與原 data concatenate 後，將其 z-normalization。

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：下方之數據，是採用 logistic regression model，Adagrad 演算法，並使用助教所抽取之 feature，epoch = 3000，lr = 0.05 的條件下，無/有 normalization 之結果：

	Training set accuracy	Testing set accuracy
Normalization 前	0.795860	0.783428
Normalization 後	0.853567	0.852957

由上方的數據可知，在相同的訓練次數、learning rate 等初始條件下，feature normalization 可大幅的提升 model 的準確度。(在沒有使用 feature normalization 的情況下，即使使用了 adagrad 演算法，仍可觀察到 model 之正確率大幅、且不穩定的跳動)

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：使用 adagrad 演算法，epoch = 3000，lr = 0.05，採用助教提供之 feature，根據不同的 λ 值，得到不同的 accuracy：

λ 值	Training set accuracy	Testing set accuracy
λ = 0.1	0.639077	0.634604
λ = 0.01	0.728171	0.719673
λ = 0.001	0.853475	0.852097
λ = 0.0001	0.853597	0.852834
λ = 0	0.853567	0.852957

以上資料顯示，再 λ 值夠小時，regularization 對 model 準確率的結果影響不大；相反的，當 λ 值較大時，越大的 λ 值會使 model 的準確率越差。

5.請討論你認為哪個 attribute 對結果影響最大？

答：使用 adagrad 演算法，epoch = 3000，lr = 0.05，採用助教提供之 feature，每次拔除一種 feature，得到以下數據

移除之 feature	Training set accuracy	Testing set accuracy
None	0.853567	0.852957
Age	0.851355	0.851790
Fnlwgt	0.852768	0.852281
Sex	0.853014	0.852527
Capital gain	***0.838334***	***0.837970***
Capital loss	0.851018	0.850193
Hour per week	0.851202	0.851360
Workclass	0.851878	0.850316
Education num	0.853505	0.852650
Education	***0.845336***	***0.845771***
Marital status	0.853874	0.852650
Occupation	***0.848069***	***0.846753***
Relationship	0.852891	0.852281
Race	0.853290	0.853080
Native country	0.852154	0.852957

從以上的數據分析，Capital gain 對於結果(accuracy)影響最大，在拔除了這項 feature 後，model 的準確率下降了將近 1.5 個百分比，除此之外，Education 與 Occupation 對 model 的準確率也有一定程度的影響(分別為 0.8%與 0.5%)。