學號: R06921095 系級: 電機所碩一 姓名: 吳睿哲

Q1:請比較你實作的 generative model、logistic regression 的準確率，何者較佳？

A1:最後結果為 logistic regression 表現較為優秀。

Generative 的預測方法為假設此訓練資料為 Gaussian Distribution,其實本身就隱含 bias 在,所以如果再 tune 過多的參數,最後的 performance 不一定在 public set 會 imporove 多少,最後還可能會在 testing data 有 overfitting 的問題。

	generative mode	logistic regression
實作說明	假設訓練資料分別由 Gaussian Distribution 產出，然後經由 maximum likelihood 去逼近此結果。 計算出 mean 與 covariance 後，可以由 Gaussian Distribution 計算出機率，而得到 distribution 的近似值，就可以得到 binary class 的機率來進行分類。	隨機初始化 w 跟 b，由 training data，經 sigmoid function 去計預測 y。 再經由 loss function 去計算 cross entropy，還有更新 gradient 經由更新的 gradient 不斷更新 w 跟 b。
準確率(kaggle)	0.84619	0.85909

Q2: 請說明你實作的 best model，其訓練方式和準確率為何？

A2 一開始我先把所有 features 跟收入的 correlation 都找出來,雖然 age 跟收入的 correlation coefficient 只有 0.22,但是跟其他 correlation coefficient 較高的 features 加入二次方中,結果卻意外的更好,得到 0.85712 的準確率,後來我又把性別跟每周工作小時的工作時數的平方考慮進去,得了更搞的準確率,我有試過用 xgboost,可以讓準確率到達 0.87 左右,但是不能使用這個套件,所以我最好的一次是使用 age,sex,capita_gain,capital_loss,跟 hours_per_week 的平方得到 0.85909 的準確率 (logistic regression)。

Q3: 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。

A3:在這次的作業中,normalization 有決定性的影響, 我的 normalize 的方法是將是將每筆資料做平均(mean)，然後算標準差(std)，以防分母為零，所以我將 std+1e-20。結果顯示 normalization 具有使預測準確率上升的效果，也較不容易

產生 gradient 爆炸的問題。一開始沒有用 normalization 時,準確率只有大概 0.80xxx 左右,做完 normalization 之後,準確率可以達到 0.84xxx,雖然還不是最好的一次,但是 performance 已經有顯著性的提升。

Q4: 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。

A4: 調整 regularization 的參數 lambda 值，以同樣的參數與 model 進行訓練的準確率比較。根據結果，可得知 lambda 值在這樣的模型下不可過高，否則無法使參數回歸至使 cost 最低的結果，此外 regularization 對預測準確率的提升效果有限。在某些訓練 epoch 較大量的 model 中可能可以抑制 overfitting 的問題，但由於此 model 中的預測率收斂速度快，導致效果有限。

Q5: 請討論你認為哪個 attribute 對結果影響最大？

A5:在找哪個 attribute 對 performance 影響最大時,我有試過找每一個 features 跟收入之間的 correlation coefficient,最高的是第 43 個 features,第 68 個 features 則是次高,但在這兩個 features 加二次項之後,performance 並沒有 age 加二次方來的更好,雖人我覺得蠻疑惑的,但如果以 performance 來看的話,我覺得 age 對結果的影響最大,再來是 sex,再來才是 work_hour_per_week。