學號：R05522621 系級：機械碩二 姓名：李哲銘

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

答：由以下的比較表可知我實作的 logistic regression 的準確率較 generative model 佳，而 logistic regression 中的梯度下降法是使用 adagrad 的方法。

Model	Public	Private
generative model	0.84594	0.84277
logistic regression	0.85454	0.85149

2. 請說明你實作的 best model，其訓練方式和準確率為何？

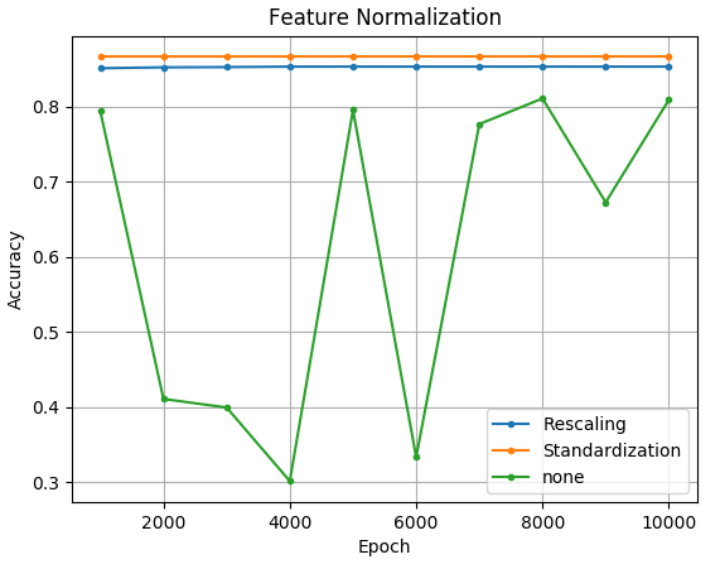
答：我實作的 best model 是使用 sklearn.svm 中的 SVC，屬於 SVM 中的方法。SVM 訓練演算法為非機率二元線性分類器。SVM 模型是將訓練資料表示為空間中的點，這樣對映就使得單獨類別的訓練資料被儘可能寬的明顯的間隔分開。

Model	Public	Private
best model (SVM)	0.85540	0.85186

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：我實作的特徵標準化方法為 Rescaling: $X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$ 及 Standardization: $X' = \frac{X - \mu}{\sigma}$，其中 μ 為 feature vector 的平均，σ 為 standard deviation。下表為標準化之準確率比較，下圖為標準化之訓練過程比較。由下圖可知標準化之後的訓練路徑變動小也平順非常多，且由下表可知標準化後的準確率表現也較好。

Feature Normalization	Public	Private
None	0.77506	0.77656
Rescaling	0.85405	0.85173
Standardization	0.85356	0.85026



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：下表為實作 logistic regression 之不同正規化(regularization)參數的準確率比較。由下表可知正規化參數小於 1 時的影響並不明顯，但當參數較大時準確率便慢慢下降。

λ	Public	Private
0(no regularization)	0.85565	0.85063
0.0001	0.85454	0.85038
0.001	0.85565	0.85075
0.01	0.85479	0.85050
0.1	0.85380	0.85100
1	0.85417	0.85050
10	0.85503	0.84854
100	0.84570	0.84031
1000	0.78501	0.78540

5. 請討論你認為哪個 attribute 對結果影響最大？

答：下表為捨棄不同 attribute 對準確率之影響。由下表可發現訓練資料捨棄 capital_gain 之後的準確率下降最明顯，因此我認為 capital_gain 對結果的影響最大。

Drop Attribute	Public	Private
age	0.85233	0.85087
fnlwgt	0.85282	0.85136
sex	0.85491	0.85112
capital_gain	0.83918	0.83564
capital_loss	0.85098	0.84793
hours_per_week	0.85343	0.84915
workclass	0.85122	0.85063
education	0.84692	0.83834
marital_status	0.85307	0.85087
occupation	0.84778	0.84621
relationship	0.85638	0.84891
race	0.85356	0.85124
native_country	0.85417	0.85100