學號：B03705028 系級： 資管三 姓名：陳星宇

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

使用助教提供的 X_train.csv，在 106 個 feature 中，使用 index 為 [0, 3, 4, 5, 6, 10, 15, 20, 21, 24, 25, 26, 27, 29, 33, 35, 41, 42, 45, 47, 49, 50, 58] 的 feature 作訓練，並假設它們的分布皆為 Gaussian distribution，以 maximum likelihood 為目標，並使用投影片中的公式計算 mean 與 covariance matrix。在本地的 training accuracy 為 0.83738，在 Kaggle 的 private testing accuracy 為 0.84203。

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

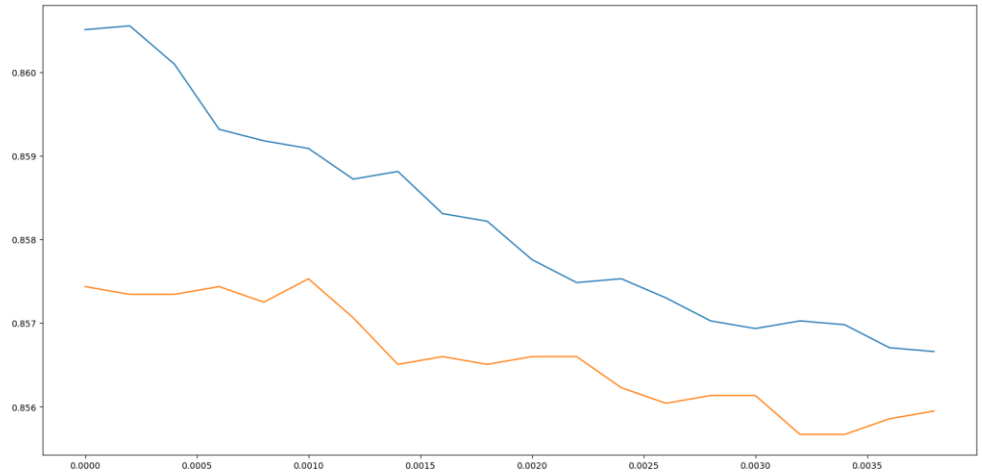
使用助教提供的 X_train.csv，除了原有的 106 個 feature，再另外加入 5 個 feature：age 的平方項、fnlwgt 取自然對數、capital_gain 開根號、capital_loss 開根號、hours_per_week 的平方項。訓練方式的部分，以投影片中提到的 cross entropy 為 loss function，並實作 ada_grad。在本地的 training accuracy 為 0.86014，在 Kaggle 的 private testing accuracy 為 0.85739。

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：

	本地 training accuracy	Kaggle private testing accuracy
實作特徵標準化前	0.84269	0.84117
實作特徵標準化後	0.86014	0.85739

由表中數據可以發現，實作特徵標準化後，無論是本地的 training accuracy 或是 Kaggle 上的 testing accuracy 均有顯著改善，推測是原本 feature 之間的 scale 相差太大 (例如 flnwgt 與 age)，導致 scale 較大的 feature 對模型的影響過大。

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：



我將 X_train.csv 分成三份，一份作為 testing data，剩下的兩份作為 training data，並使用第 2.題提到的 discriminative model 作為訓練模型。圖中的縱軸代表 accuracy，橫軸代表 lambda 的值，藍線代表 training data，黃線代表 testing data。由圖中可知，隨著 lambda 的值越來越大，training 跟 testing 的 accuracy 都呈現下降的趨勢，推測可能是這個 model 還沒有 overfitting，或是應該要使用 l1 之外的 regularization method。

5.請討論你認為哪個 attribute 對結果影響最大？

我先計算第 1.題的 generative model 初步篩選出的 23 個 feature 各自對 dependent variable 的相關係數，再訓練模型 23 次，一次取出一個 feature，以觀察少掉那個 feature 之後，對模型準確率的影響。最後的結果，index 為 33 的 Married-civ-spouse 有最高的相關係數(0.44469)，而且少考慮 Married-civ-spouse 的模型的準確率也最低，為 0.82076，相較於考慮 23 個 feature 的模型，準確率整整下降了 0.1671。因此，我認為 Married-civ-spouse 對結果的影響最大。