學號：B04901117 系級： 電機三 姓名：毛弘仁

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

答：

Logistic regression 的準確度較佳。

Generative model

我的 continuous features 用 multivariate Gaussian distribution 去 model，discrete features 用 independent Bernoulli distribution，也就是假設每個 discrete feature 都互相獨立，然而這樣的假設很明顯是錯誤的，e.g. Peru == 1 就代表 France == 0。

不過 accuracy 還算不錯：當>=50K 及 <50K 的分佈使用不同的 covariance matrix 時，得到的準確率為 0.835；使用相同的 covariance matrix，準確率 0.78230。可見在這項 task 當中，讓分隔兩類的 hyperplane 複雜度高一點可能比較好。

Logistic regression

我做了 continuous input features 的 normalization、每個 epoch training data 順序的 randomization，使用的 batch size = 800，epochs = 10。Validation set 在每次要開始 train 的時候，會隨機挑選出 training data 的 20%。使用 exponential decay 作為 optimizer（效果好像不輸 Adagrad 且速度較快），初始 learn rate 是 0.01，decay rate = 0.95。

Validation accuracy 是 0.852。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我利用 Keras 做出 hidden layer size = 120 且用 ReLU activation 的 feedforward model，使用 Adam optimizer，達到 0.858 的準確率。網路上查到一般的 task 用一個 hidden layer 就足夠，hidden layer size 設多少則有好幾種 rule of thumb 可以參考，我挑了其中一種，讓 size 介於 input size 和 output size 之間。

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：

沒有做 normalization 時，validation accuracy 會在 0.779 和 0.24 中間跳動，十分不穩定，推測是因為 output 全部是 0 或 1。

有做 normalization 時，validation accuracy 的 0.852。

由此可知，normalization 對於 weights 的穩定性是重要的，可提升 accuracy。

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：

沒有使用 regularization 時，validation accuracy = 0.852。

使用 l2 norm 後，lambda = 0.1，validation accuracy = 0.653；lambda = 0.01，validation accuracy = 0.795，很穩定；lambda = 0.001，validation accuracy = 0.8。

從實驗結果看，regularization 似乎在這項 task 沒有太大的作用，而且如同李宏毅老師所言，若要避免 overfitting，early stopping 已經是好方法。

5.請討論你認為哪個 attribute 對結果影響最大？

經過實驗，我發現 capital_gain 欄位若去除掉，對於 accuracy 的影響是高於移除其他 attributes 的。這對本次的 task 合理，因為有錢人的資本拿去賺更多錢，是常有的事。雖然我不是有錢人啦。