學號：R05943139 系級：電子碩一 姓名：張育瑄

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

利用上課教的公式：

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

先求得 $P(C_1)$ 和 $P(C_2)$，然後再利用 Guassian Distribution 及 Maximum Likelihood 求出 $P(x|C_1)$，便可以得到我們最終想要的 $P(C_1|x)$。

我訓練出來的結果在 kaggle 上準確率為 78.13%。

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

使用所有的 feature 除了 fnlwgt 做 training，把 continuous 的 feature 加到 2 次方，並做 normalization，最後再加上 regularization 來避免 overfitting，λ 設為 100。

在 kaggle 上的準確率為 85.55%。

另外，若加到 3 次方，雖然在 kaggle 上的分數非常高，但在 validation set 上的表現比 2 次方差，有 overfitting 的可能。

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：

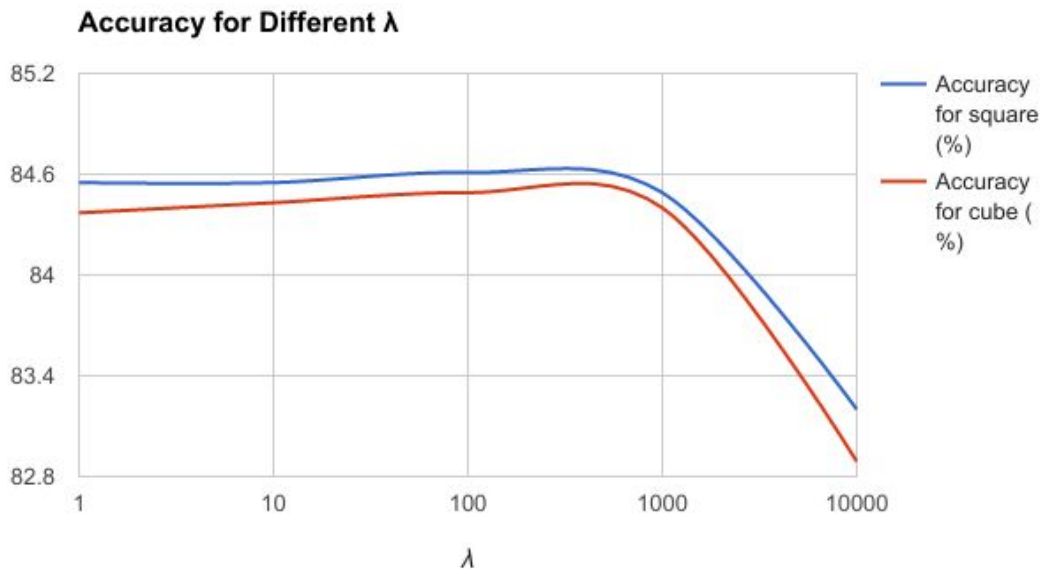
若不做 normalization，cross entropy 便會上下震盪，無法穩定收斂，根據觀察，我推測是因為算 sigmoid 的過程中，e 的指數運算會 overflow，雖然我有做 clip 的動作來防止 nan 的產生，但 clip 可能就是使誤差變大的元兇，因此最後結果很差。

Accuracy
with normalization：84.43%
without normalization: 80.59%

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：

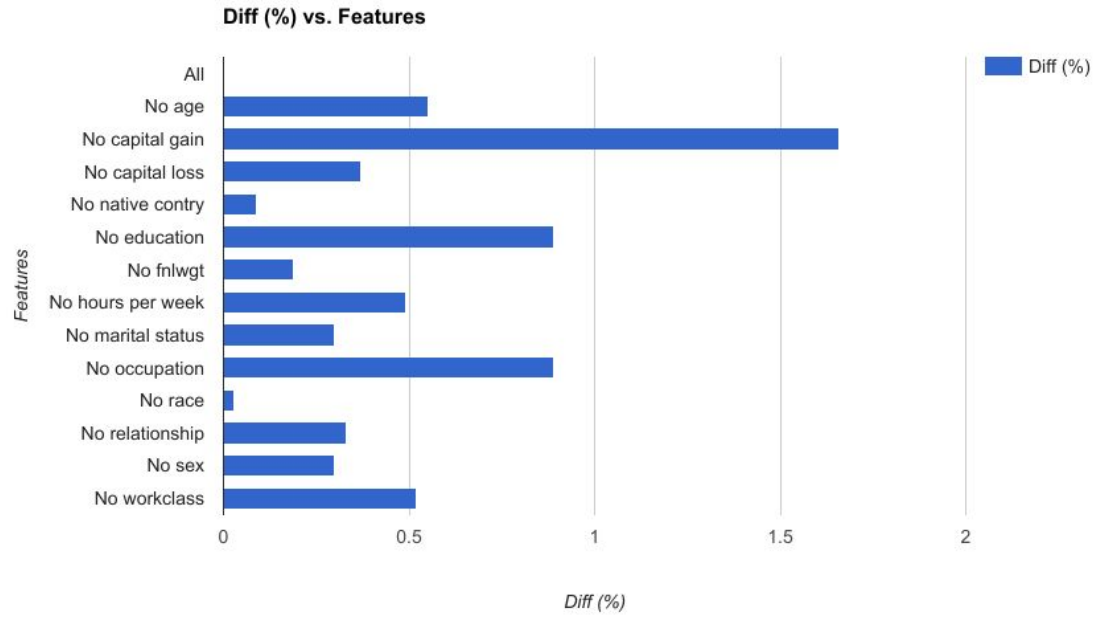
以下表格為我在上述 logistic regression 的方法下嘗試用不同次方以及 λ做 regularization 的結果，accuracy 是在我的 validation set 上的準確率。

λ	Accuracy for x^2 (%)	Accuracy for x^3 (%)
1	84.55	84.37
10	84.55	84.43
100	84.61	84.49
1000	84.49	84.40
10000	83.20	82.89



5.請討論你認為哪個attribute對結果影響最大？

先使用所有 feature，continuous 的 feature 都只有 1 次方來做 logistic regression，再依序分別將一種 feature 拿掉，最後觀察每一種結果跟用全部 feature train 出來的結果差異(Diff)，以下圖表即為實驗結果。



由圖表便可清楚地看到，對預測結果影響最大的前五名 feature 為：

1. Capital Gain
2. Education
3. Occupation
4. Age
5. Workclass