Homework 2 Report - Income Prediction

學號：b05611038　系級：生機二　姓名：張育堂

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

	generative model	logistic regression
正確率	0.652595	0.8524

logistic regression 的正確性比較高，我想是因為在只有 3 萬多筆 training data 的狀況下，單純貝氏機率運算所生成的 generative model 是非常不準確的，但是資料量增加 100 倍，也許 generative model 就可以到很高的精確性。

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

我的 best model 是使用 momentum 再加上 Adagrad 去寫的，只使用單維 feature，但是非 binary feature 都有作標準分數之後，再將所有的數值縮小到 1 以下以利指數運算。

3. (1%) 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。(有關 normalization 請參考：https://goo.gl/XBM3aE)

我於模型內實作的是標準分數，也就是 $\frac{x-\bar{x}}{\sigma}$ 的形式，我這些特徵縮放的步驟，對於權重的運算是很有幫助的，因為這些運算出來的新值，更能代表每個點在同一個 feature 的狀況和程度，而不會被數字大小被蒙騙。

4. (1%) 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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

在高維 feature 中，正規化才會比較常出現，但說實在話，這並不是一個非常適合 machine learning 的方式，因為其實在高維 feature 同時放在矩陣中，它其實也代表一個線性的新的 feature，在 gradient descent 中，若是不必要的參數，其實它的 weight 也會越來越小，因此對於 feature 的正規畫對於 model 的幫助並不大甚至有時候還會對數據失真。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

Print 出最終 weight 之後，發現 capital gain 跟 capital loss 雖然代表的意義對於收入來說是相反的，但是實際上他們的權重都是正很多的，我想原因是因為本身會進行投資進而擁有這兩項數據的人，他們本身的收入較高是正常的現象；weight 中，負的最多的是從事勞力工作的人，我想這也很符合現實狀況，這些低階層的勞力工作者要有高收入本來就是一件不容易的事。在這次 logistic regression 中程式所學到的參數的確也非常符合實際情況，這也讓我對這門課更加佩服。