學號：B03705027 系級： 資管三 姓名：鄭從德

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

我的 generative model 就是照著老師投影片上的公式刻的：在算出 mu0 mu1 之後算出個別的 sigma 矩陣，然後就可以直接算出 w 和 b。我的 training set 是採用 X_train.csv(也就是助教切好的)中全部的資料點，因此也就只有上傳 kaggle 的一次成績，正確率為 0.84128。

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

我的 discriminative model 跟 hw1 十分的像，採用的是 SGD 加上 adagrad 的優化，跟 hw1 基本上只差在 sigmoid function 而已。訓練方式是使用所有的 training data，經過標準化，但沒有使用 mini batch，所以一個 iteration 是將所有的 data point 跑過一次。我做出 kaggle 上 public 最佳的成績約是 learning rate 為 0.017，iteration 為 800 的時候，正確率為 0.8534。其他次的成績大約都在 0.845 到 0.853 之間。至於關於正規化的影響會在第四題討論。

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。

加入 feature normalization 前後我的準確度都有非常明顯的提升，在做標準化之前，正確率大多是 0.81 到 0.83 之間(Testing Accuracy)，做了標準化之後，正確率大概直接提升了 2%以上。而且這不只是是用在 logistic regression，在我的 neural network 以及用 keras 測試的結果都是標準化後會上升好幾個百分點，尤其 neural network 的 training accuracy 在標準化之後可以提升到接近 90%，keras 甚至可以超過 90%。雖然 training accuracy 並沒有太大意義，且 90%的 accuracy 多半已經是 overfit，但是還是可以解釋為 normalization 之後的 data 比較有辦法拿來解釋這個分類問題，也可以 train 出較好的 model。

至於 generative model，跑出來的結果是幾乎相同的。

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。

這次的 logistic regression 跟作業一還有一個比較大的不同就是關於 lambda 的選擇。因為這次的 features 都經過正規化，值較小，所以 lambda 的值也要選得較小。我一開始把 lambda 設成 2，會完全 train 不動，就連 training accuracy 都到不了 80%。

但是在 Lambda 下降之後就有好轉；大約下降到 0.05 左右就會對於 training accuracy 有稍微下降(約從 86.5%下降到 85.5%左右)，而 testing accuracy 上升的效果。
以下提供一筆做正規化前後 kaggle 成績上的變化
0.8361 -> 0.8411
除了 logistic regression 以外，在我實作的 neural network 中也加入了 regularization。因為沒有使用 SGD，所以在正規化方面 lambda 要選擇比較大的數。實作結果約選擇 lambda = 0.5 到 1.2 之間效果最佳，training accuracy 會從 89% 減到 86%，十分有效的減少 overfitting 的現象。

5.請討論你認為哪個 attribute 對結果影響最大？

Continuous variable 跟 categorical variable 要分開探討。
我們先將所有資料經過標準化，將所有的特徵都除以該特徵的最大值，將最大的數變為 1。為什麼要使用這種方式標準化是為了怕非連續特徵的加權發生錯誤，我們這種標準化方式可以讓所有的 1 , 0 保持不變。

Continuous variables 跑出 w 後，乘上每個特徵的平均值(消除單位不同的影響)，看看哪一個特徵貢獻最多值到 z。結果是 hour_per_week，一周工時，工時月常，越有機會年收 50K 以上。第二高的是 age，年紀越大越有機會年收 50K 以上。

Categorical variable 因為最後都會乘上 0 或 1，所以我們應該直接比較每一個 w 值，因為乘上他們的平均數是沒有意義的。在結果中我們應該比較同一問題哪一個每個選項的 weight 差異最大。比較結果影響最大的是學歷，如果有博士學位(Doctorate)會先為 Z 值貢獻 2.21(同時這也是純粹比 w 最高的一個值)，而只有 preschool 的會讓 z 值-1.56。