ML2017 HW2 Report

學號:B02901124 系級: 電機四 姓名:黃柏翔

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

將所有的資料分成兩群：>50K 及<=50K 的人，先計算兩群各自的人數並計算 106 個特徵在兩群中各自的平均值(mean)，並藉由公式計算兩群各自的 sigma 值，便能得到假定這兩群人爲高斯分佈的參數。接著計算 shared sigma。最後在計算機率時利用投影片上的結論去定義 weight 及 bias 便能作出 prediction。

最後我的 Generative model 的正確率在 0.841~0.846 之間。

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$z = \underbrace{(\mu^1 - \mu^2)^T \Sigma^{-1}}_{w^T} x \underbrace{- \frac{1}{2}(\mu^1)^T \Sigma^{-1} \mu^1 + \frac{1}{2}(\mu^2)^T \Sigma^{-1} \mu^2 + ln \frac{N_1}{N_2}}_{b}$$

```
predict(X_test,m1,m0,sigma,n1,n0):
print "Predicting..."
sig_inv = np.linalg.inv(sigma)
x = X_test.T
w = np.dot((m1-m0),sig_inv)
b = (-0.5) * np.dot(np.dot([m1], sig_inv), m1) + (0.5) * np.dot(np.dot([m0], sig_inv), m0) + np.log(float(n1)/n0)
a = np.dot(w, x) + b
y = sigmoid(a)
out = []
```

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

首先我直接利用助教取好的 feature 做 training。所有的 feature 都用。而我在第 0,1,3,4,5 項的 feature 有先進行 normalization 的動作。在 training 時是 32561 筆資料讀完後一次更新所有的 106 個 weight，並且中間有實作 adam 以提升準確率。

最後我的 Discriminative model 正確率在 0.8525~0.8545 之間。

```
for i in range(itera):
    print "Iteration: %d     " % (i+1),
    pred = expit(np.dot(X,w))
    diffy = np.subtract(pred,Y)
    loss = np.sum(np.negative(Y*np.log(pred)+(1-Y)*(np.log(1-pred))))
    #print "Loss: %f" % float(loss)
    g = np.dot(X.T, diffy) * 2 + lamda*2*w
    m = beta_1*m + (1 - beta_1)*g
    vec = beta_2*vec + (1 - beta_2)*(g**2)

    m_hat = m/(1 - beta_1**itera)
    vec_hat = vec/(1 - beta_2**itera)
    w = w - a*m_hat/(np.sqrt(vec_hat) + epsilon)
    print "Loss: %f\r" % float(loss),
print "Final Loss: %f                " % float(loss)
```

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。**

在沒有實作 normalization 之前，一些比較數字範圍比較廣或比較大的特徵很容易 dominate 整個 prediction 的方向。有時候會使 sigmoid function 預測出的值偏向單一的預測值。加了 normalization 後各項特徵的 weight 的更新比較均衡，準確率從 0.81 上升到 0.85 左右。

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。

在 106 維度的特徵中，regularization 的影響似乎不是那麼大，我將 lamda 從[0 0.1 0.01 ... 0.00001]做調整似乎並沒有顯著的效果。

5. **請討論你認為哪個 attribute 對結果影響最大？**

由於我一開始讀檔案時是自己去讀取 train.csv 檔，後來才用助教的 feature。在一開始時有去選取 feature 時並有觀察選取的 feature 與正確率的關係，因此現在利用那時的數據做分析。

首先我是將有數字的部分取出，也就是 age, fnlwgt, education_num, capital_gain, capital_loss, hours_per_week 這幾項，正確率約爲 0.79 左右。

而後我判斷性別,種族,及國籍一定有很大的影響，因此在這三項有去賦予一些值加入 training，正確率大約上升到 0.81 左右。從這小小的實驗我判斷這幾項與正確率都有顯著的關係，但要說那個影響最大的話，我猜測在性別，種族或國籍中有決定性的因素。

但在最後的實驗中，我每次逐一將一兩個 Feature 刪去去求得最後的正確率，發現在刪去了 capital_gain, capital_loss 這兩項時，最後 accuracy 下降程度達到 2%，其他的 Feature 頂多下降 1%。所以推斷這兩項對於結果的影響最大。