HW2

學號：R05943005　系級：　電子所碩一　姓名：呂丞勛

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

答：

定義年收入<50k 的為 C0，>=50k 的為 C1。假設 C0 與 C1 的 data 是從一個 Gaussian distribution 所取樣出來，則可以推算該兩個 class 的平均值與標準差，又再假設兩個群體的標準差相同，則可以使用下式來計算 Test data 屬於 C0 的機率:

$$\mathrm{P(C0|x)} = \frac{P(x|C0)P(C0)}{P(x|C0)P(C0) + P(x|C1)P(C1)}$$

若大於 0.5 則將該次 test data 區分到 C0，否則區分到 C1，由此方式計算的結果，丟到 kaggle 上的測試準確率為 83.771%。

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

答：

實作 discriminative model 時，並非使用 sigmoid 來作為最終 output，我使用等效的 softmax 來實作(output 有兩個 node)，利用微分來作 gradient descent，對於每個 weight 來說，每次訓練得到的 graident 為 $(\sum_{i=0}^{1}(\mathrm{y_i} - t_i)w)$，$\mathrm{y_i}$表示屬於哪個 class 的機率，learning rate 的部分使用 RMSProp 來讓 model 可以訓練的快一些，該訓練方式(不含 regularization)，為 85.356%。

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：

以下分成幾種狀況說明。

a. 完全沒有 normalization，當 feature 沒有先做 normalization 直接下去訓練模型時，該情況下，training loss (使用 cross entropy)完全沒有下降，並且會有很大幅度的震盪，由於第二個 attribute 的數值都相當大，會造成對應到的 weight gradient 很大，使得 weight 難以達到穩定狀態。

b. 對所有 feature 都 normalization，此情況下，相較於第一個狀況 weight 較容易達到穩定狀態，因而可以獲得不錯的正確率 (85.356%)。

c. 只對 continuous 的 feature 做 normalization，由於對 discrete 的 feature 做 normalization 的話，會使得本來為 0 的數值變成>0，這樣跟 1-hot 的 encoding 方式有矛盾，因此只對 continuous 做 normalization，果然可以得到比第二個更好的正確率 (85.614%)。

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：

Regularization Coefficient	Testing Error (Kaggle public)
1	0.85233
0.1	0.85369
0.01	0.85344
0.001	0.85356
0.0001	0.85332
0	0.85356

(此題利用 discriminative model 討論)Regularization 對於 model 而言，是用來避免 overfitting 於 training data，但過大的 regularization coefficient 反而會造成 test 的結果不甚理想，由上表的測試可以觀察出這個趨勢，當 regularization 到達 1 實，testing accuracy 明顯的比其他結果都還要差，而有作些許 regularization 的結果反而較好。

5.請討論你認為哪個 attribute 對結果影響最大？

對於薪水而言，會認為教育程度的結果影響最大，不過這部分由於 feature 抽出來的是 discrete，因此並沒有真實下去測試對於準確度的影響。但是 continuous 的 feature 的部分，認為每周工作時數影響不小，在訓練的時候，有將連續數值的資料的二次項加入增加 feature 的類別，可以發現正確率會有上升許多的現象，從 85.3%上升到 85.6%，因此認為對於結果影響，continuous 的部分也佔相當大的成分。