Machine Learning 2017 Spring
Homework 2 Report

學號：B03902048
系級：資工三
姓名：林義聖

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

   答：我先將全部的訓練資料透過 label 分成兩筆，分別計算各自的 mean 和 covariance matrix，在套用下方公式算出 $P(x|C_1)$ 與 $P(x|C_2)$

   $$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T(\Sigma)^{-1}(x-\mu)\right\}$$

   然而，在過程中，我卻發現 overflow 導致的種種問題很難解。於是我改用另一個 covariance matrix：

   $$\Sigma^* = P(C_1)\Sigma^1 + P(C_2)\Sigma^2$$

   接著，套用助教投影片上提供的公式，直接算出 $P(C_2|x)$，也就是 $y = 1$ 的機率。我的 generative model 可以在 public set 上得到 84.165% 的準確率。

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

   答：首先，將所有特徵的三次方項接上去，並接上一整個欄位的 1 作爲 bias，再將特徵正規化 (z-score normalization)。接著，切分出 6000 筆資料的 valid set 後開始做 gradient descent。起初，我手動設置 learning rate，發現訓練成果很差，於是改用 adagrad 在過程中調整 learning rate。而每經過 100 個 epoch，我就預測一次 valid set，藉由觀察在 valid set 上的 accuracy 和 loss，我可以得知何時該停止。

#	**learning rate**	**iterations**	**train accuracy**	**valid accuracy**
1	0.05	5000	85.70%	85.65%
2	0.1	1500	85.72%	85.63%
3	0.5	400	85.70%	85.63%
4	1	5000	85.70%	85.62%

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：尚未進行特徵標準化以前，由於特徵中的某些數值特別的大，以致在做 sigmoid 時，經過指數函數轉換以後很容易溢位。在溢位發生後，由於參數變動過大，甚至導致我的模型訓練不起來。而加入標準化後，則可以大幅改善這種情況。

   我一共實作了三種標準化方法，而以下結果，皆是以 1000 個 epoch、6000 筆 valid data 和一次方加三次方項的參數，來做實驗而得：

#	method	train accuracy	valid accuracy
1	none	76.98%	76.60%
2	min-max [0, 1]	85.32%	85.36%
3	min-max [-0.5, 0.5]	83.86%	83.62%
4	z-score	85.70%	85.62%

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：根據實驗得到的結果，可以看出不論是 L1 regularization 還是 L2 regularization，都能一定程度地提升模型準確率。在訓練回合數提升的前提下，valid loss 在有實作正規化的模型中，能夠持續地下降更長的時間。因此我們能確知，有正規化的模型，確實能避免過度擬合 training set，而能夠在 valid accuracy 上取得更好的成果。

#	method	lambda	iterations	train accuracy	valid accuracy
1	none	none	400	85.70%	85.62%
2	L1	0.01	1600	85.54%	85.78%
3	L1	0.001	500	85.75%	85.62%
4	L2	0.005	550	85.73%	85.62%
5	L2	0.001	500	85.74%	85.67%

5. 請討論你認爲哪個 attribute 對結果影響最大？

   答：我試著一次移除一整組的特徵，並訓練 500 回合之後，觀察 valid accuracy 來評估每個 attribute 對模型預測準確率所造成的影響。

#	removed features	valid accuracy
1	none	85.13%
2	age	85.40%
3	fnlwgt	85.18%
4	sex	85.17%
5	capital gain	83.82%
6	capital loss	85.18%
7	hours per week	85.37%
8	workclass	85.23%
9	education	84.37%
10	marital status	85.18%
11	occupation	84.93%
12	relationship	85.37%
13	race	85.13%
14	native country	85.00%

   所以去除一些準確率上微小幅度的變化不看，還是可以明顯看出 capital gain 對結果影響最大。事實上，capital gain 就是從一些資本商品：股票、債券、房地產等的交易所取得的收益。因此，合理推斷 capital gain 必然直接也間接地影響一個人的年收入。