機器學習作業二 report

學號：B05901040　系級：電機三　姓名：蔡松達

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

此處使用 MinMaxScaling 進行實作，實作時只針對 0,1,3,4,5 進行標準化，logistic regression 之設定參數為 w 的 learning rate = 0.5，b 的 learning rate = 0.5，epochs = 35000，adagrad，而兩者使用的 data 皆是經處理過的 X_train、Y_train 及 X_test。

Model	Public accuracy	Private accuracy	Average accuracy
generative model	0.84545	0.84105	0.84325
logistic regression	0.84840	0.84559	0.84700

基本上由於 logistic regression 可以調整的參數較多，generative model 使用的是既定假設的高斯分布進行實作，因此可變動性較 logistic regression 為低，透過調整 learning rate 與 epochs 之值往往可以得到較好的預測結果。（不過也可以透過改變分布模型對於 generative model 進行更動，此處只考慮使用高斯分布實作 generative model，因此可變動性較 logistic regression 為低。）

2. 請說明你實作的 best model，其訓練方式和準確率為何？

此處使用 sklearn.ensemble 中的 Gradient Boosting Classifier 套件來實作，得到結果如下表所示。

Model	Public accuracy	Private accuracy	Average accuracy
Gradient Boosting Classifier	0.87297	0.86954	0.87126

基本概念為利用前一輪迭代弱學習器的誤差率來更新訓練集的權重，使用了前向分佈演算法。在 Gradient Boosting Classifier 的迭代中，假設前一輪迭代得到的強學習器是 ft−1(x)，損失函式是 L(y,ft−1(x))，本輪迭代的目標是找到一個 CART 迴歸樹模型的弱學習器 ht(x)，讓本輪的損失 L(y,ft(x))=L(y,ft−1(x) ht(x))最小。本輪迭代找到決策樹，要讓樣本的損失儘量變得更小。

此處使用的參數皆為套件中的預設值，重要參數包含：

I. n_estimators: 預設是 100，最大的弱學習器的個數，或者弱學習器的最大迭代次數。

II. learning_rate: 預設為 0.1。

III. loss: 使用對數似然損失函式「deviance」。

Ref: https://goo.gl/kkUX6K

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。

此處使用 MinMaxScaling 進行實作，logistic regression 之設定參數為 w 的 learning rate = 0.5，b 的 learning rate = 0.5，epochs = 35000，adagrad，與沒有進行 normalization 處理的 accuracy 做比較。

未使用 feature scaling 的 Model	Public accuracy	Private accuracy	Average accuracy
generative model	0.84570	0.84129	0.84350
logistic regression	0.67051	0.67055	0.67053
有使用 feature scaling 的 Model	Public accuracy	Private accuracy	Average accuracy
generative model	0.84545	0.84105	0.84325
logistic regression	0.84840	0.84559	0.84700

由上表可見 generative model 的 performance 與有無進行標準化較無關聯，結果相當接近，可能原因為其模型已經受到限制，因此進行 scaling 與否影響較小（進行計算過程中就會有類似 normalize 的效果）；而若是在 logistic regression 情況下若未使用標準化，則會得到相當差的結果，主要原因為資料中的 feature 其中幾個 continuous 的值相當大，其他 discrete 的資料又因為進行 one-hot encoding 而分散其資訊，導致進行 gradient descent 時將嚴重失去平衡，gradient 數值容易爆炸，數值大小的 scale 過大的情況下導致結果相當差。

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。

此處使用 MinMaxScaling 進行實作，logistic regression 之設定參數為 w 的 learning rate = 0.5，b 的 learning rate = 0.5，epochs = 35000，adagrad。

Model	Public accuracy	Private accuracy	Average accuracy
$\lambda = 0.001$	0.84840	0.84559	0.84700
$\lambda = 0.01$	0.84840	0.84571	0.84706
$\lambda = 0.1$	0.84840	0.84547	0.84694
$\lambda = 1$	0.84656	0.84375	0.84516
$\lambda = 0$	0.84840	0.84559	0.84700

由結果可見加上 regularization 的效果其實並不大，較好的狀況大約落在 $\lambda = 0.01$ 附近，當 $\lambda > 0.1$ 之後結果會越來越差，原因是因為 regularization 已經過度主導了整個 gradient descent 的進行，導致較難達到好的解。

5. 請討論你認為哪個 attribute 對結果影響最大？

attribute	3	28	13	91	8	78	46	5	67	4
weight	31.9	-16.1	-9.28	-8.94	-7.61	-7.02	-4.79	2.91	-2.90	2.82

此處將使用 logistic regression 的 weight matrix（取絕對值）前十大的數值列在上方表格中，由此可見第三個 attribute 的比重最大，接著我將該 attribute 刪去後進行 logistic regression 實作，得到的 testing accuracy 從 0.84700 降為 0.83299，可見單純刪去此 attribute 便會對於 accuracy 產生影響，此 attribute 為 capital-gain（資本利得），因此與年收入高低想必也是有相當密切的關聯。