Machine Learning - Homework 2 Report

學號：B06902049 系級：資工二 姓名：林首志

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

下表是兩種model的準確率：

Model	Public Score	Private Score
Generative	0.82764	0.82115
Logistic Regression	0.85945	0.85861

可以看出Logistic Regression的準確率比較佳。

2. 請說明你實作的best model，其訓練方式和準確率為何？

   我實作的best model使用了scikit-learn的GradientBoostingClassifier（也就是Gradient Tree Boosting演算法），learning_rate設為0.05，n_estimators設為300，max_depth設為6。我使用的訓練資料是feature scaling過後的X_train和Y_train，使用scikit-learn內建的演算法訓練。其準確率如下：

Model	Public Score	Private Score
Best(Gradient Tree Boosting)	0.87604	0.87372

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

   下表為有無實做標準化的比較：

Model	Public Score	Private Score
Generative(With normalization)	0.84570	0.84092
Generative(Without normalization)	0.82764	0.82115
Logistic Regression(With normalization)	0.85945	0.85861
Logistic Regression(Without normalization)*	0.78808	0.78245
Best(With normalization)	0.87604	0.87372
Best(Without normalization)	0.87604	0.87372

   *Logistic Regression如果不做normalization，在訓練時無法收斂

   根據上表，我們可以發現對於Logistic Regression和Generative Model，實作輸入特徵標準化都會讓模型的準確率提昇。然而對於Best model，因為是基於決策樹的模型，feature normalization並不會影響結果。

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。

   下表是有無實做L2 regularization(不對bias項做regularize)的結果：

Model	Public Score	Private Score
Logistic Regression(With regularization, $\lambda = 0.1$)	0.85749	0.85640
Logistic Regression(With regularization, $\lambda = 0.01$)	0.85945	0.85910
Logistic Regression(With regularization, $\lambda = 0.001$)	0.85945	0.85861
Logistic Regression(Without regularization)	0.85933	0.85886

可以發現微量的regularization可以稍微幫助提昇準確度，但整體而言並沒有顯著的影響。

5. 請討論你認為哪個attribute 對結果影響最大？

   我對Logistic Regression訓練出來的$\mathbf{w}$向量的分量做了分析，發現capital_gain和capital_gain的平方項的分量的絕對值是最大的，因此我認為capital_gain對結果影響最大。實際測試的結果是，去掉capital_gain之後，Public Score降低為0.84484，而Private Score降低為0.84117，下降的幅度遠比去掉其他attribute還多。