Machine Learning HW2: Income Prediction

學號：R06922030　系級：資工碩一　　姓名：傅敏桓

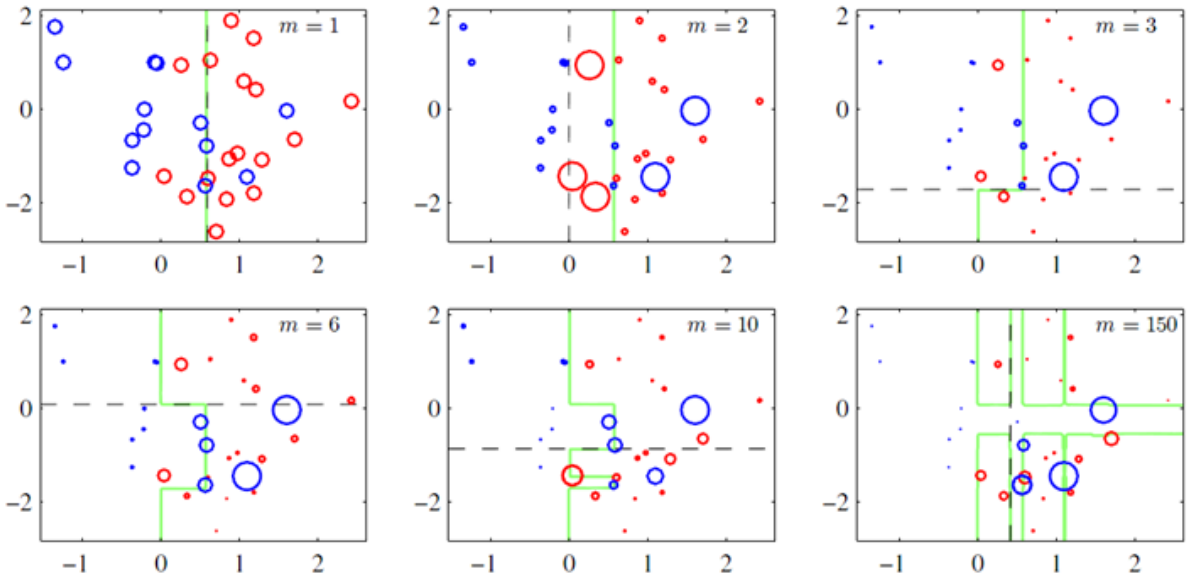
1.請比較你實作的 generative model、logistic regression 的準確率，何者較佳？

單純以這次作業在 Kaggle 的 private 分數而言，logistic regression 的表現比 generative model 來得好，兩者的 private 分數分別是 0.85124 和 0.84264。試著從兩種模型本質上的差異來探討這個問題，我們使用的 generative model 比較像是最大似然估計的方法，目標是找出一組參數使得機率模型最可能產生我們所看到的觀測值；而 logistic regression 是回歸問題，目標是找出一組參數讓模型輸出的結果和真實觀測值誤差越小越好，兩種方法的目標並不等價。

另外，generative model 是直接得到對於訓練集的最佳解，這樣的模型即便在訓練集有不錯的結果，也不一定能在測試集上面有好的表現，反而是採取逐步縮小誤差來進行參數更新的回歸模型在這方面有比較高的彈性，這也是我認為 generative model 表現比較好的主要原因。

2.請說明你實作的 best model，其訓練方式和準確率為何？

這次在 Kaggle 上得到最高分的是使用 Xgboost 得到的分類結果。Xgboost 套件實現了高效率的 Gradient Boosting 方法，其核心概念是試圖將眾多的弱分類器（簡易模型）結合成強分類器加強學習效果，而 Gradient Boosting 即透過梯度下降的方法來實作 Boosting 的方法，每一次建立模型是在之前建立模型的損失函數的梯度下降方向。下圖是 Boosting 的演示圖，每次加入新分類器時都會增加上一次分錯的數據點的權重。



(取自 http://www.cnblogs.com/LeftNotEasy/archive/2011/01/02/machine-learning-boosting-and-gradient-boosting.html)

這次嘗試 Xgboost 在分類問題上的效果，並沒有更動 Python Xgboost 套件包分類器預設的參數（使用 100 個最大深度為 3 的決策樹分類器，Boosting 學習率 0.1…），並對資料點進行標準化處理，在 Kaggle 上最後分別得到 public 分數 0.86732、private 分數 0.86549。

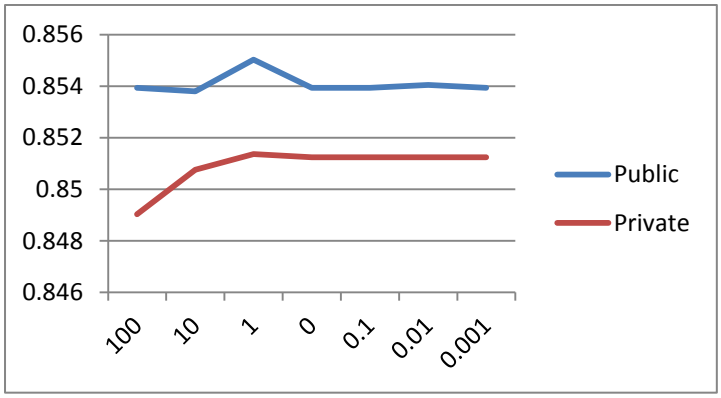
3. 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。

下表列出有無實作標準化在 private 分數上的差異。實驗設定同上述。有做特徵標準化的結果在 logistic regression 有顯著進步（應該還有部分原因是沒標準化導致 overflow 容易發生）；generative model 表現略有進步；在 Xgboost 上表現則無明顯進步，推測可能是因為 Xgboost 是採用決策樹的分類方式。

(Measure=Private score)	沒有標準化	有標準化
Logistic Regression	0.78688	0.85124
Generative Model	0.84277	0.84264
Xgboost	0.86549	0.86549

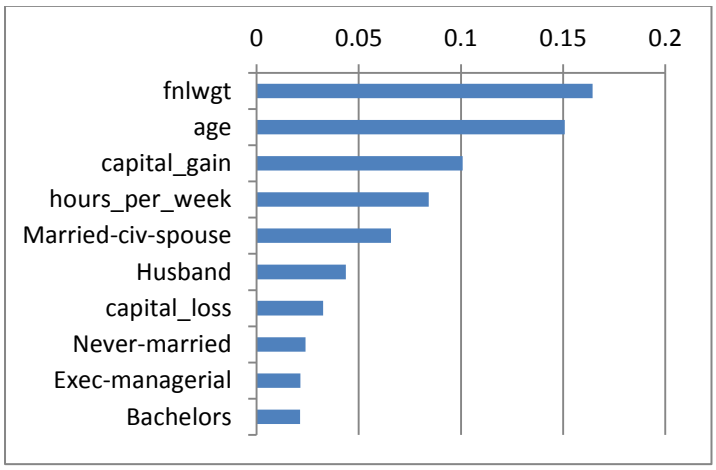
4. 請實作 logistic regression 的正規化(regularization)，並討論其對於你的模型準確率的影響。

實驗設定：隨機決定模型權重，實作 Adagrad 梯度下降，跑 1000 個 epochs，結果如下表。本次作業中 λ=1.0 的表現略好於其他值，λ< 0.1 時正規化對結果幾乎沒有影響，而在 λ> 1.0 時則反而使結果變差。



5. 請討論你認為哪個 attribute 對結果影響最大?

使用 scikit-learn 提供的 RandomForestClassifier 找出各特徵的重要度指標，可以發現這些指標也大致和現實中判斷收入水平的重要指標相符。分別嘗試只使用重要度前 5 名、前 10 名和前 20 名的特徵做分類問題，也可以得到和使用全部特徵值相近的結果。測試結果列於下表。



	Private 分數
取前 5 個特徵	0.79621
取前 10 個特徵	0.82717
取前 20 個特徵	0.84105
取全部特徵	0.85124

（以上以 logistic regression 測試，實驗設定同上述）