機器學習(Machine Learning)
Homework 02 – 2017/10/26

工工碩二 R05546030 **彭新翔**

1. 請比較你實作的 Generative Model、Logistic Regression 的準確率，何者較佳？

分別以兩種模型和作業提供資料進行訓練上傳至 Kaggle 後跑分結果如下：

	Private Score	Public Score	RMSE
Generative Model	0.84240	0.84520	0.84380
Logistic Regression	0.85603	0.86179	0.85891

如上所示，顯然以 Logistic Regression 所實現的準確率較高。

2. 請說明你實作的 Best Model，其訓練方式和準確率為何？

以不同方式訓練模型所得之結果第一題表格中所示，其中以 Logistic Regression 方法的準確率較高，為此次作業中我的 Best Model，準確率分別為 0.85603 (Private Score) 與 0.86179 (Public Score)。

在訓練時先讀入 Training Data 進行梳理並以 Logistic Regression 方法進行訓練，由於若採用固定的 Learning Rate 可能致使訓練緩慢（過小）或無法收斂（過大），在實作中採用 Adagrad 方法動態改變，初始的 Learning Rate 設置為 0.05 而迭代次數取 3000 次，並對 age、fnlwgt、capital_gain、capital_loss 及 hours_per_week 等特徵資料進行項次的擴張，其餘特徵不對其進行擴張但維持一次項。

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。

如上題中所述，實作的 Best Model 中已有對某些特徵資料進行標準化，固定其他條件改以未對特徵進行標準化所得結果比較如下所示：

	Private Score	Public Score	RMSE
Nonnormalization	0.76575	0.76732	0.76653
Normalization	0.85603	0.86179	0.85891

可見在沒有進行標準化的基礎下所得到的準確率會較低，此種結果十分顯而易見，因為不同特徵的量度標準與單位不同，甚至是分布也都不盡相同，在沒有進行標準化的狀況下，不同特徵間相比較的數值可能很大，但對整體來說其實比利並不是那麼多，但由於數值大進而使得訓練時容易被數值大的特徵所影響，最後所得到的模型準確度也會失真。

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。

同樣地，實作的 Best Model 中已是正規化（使用 $\lambda = 0.01$）之後所得到的結果，固定其他條件並改以不進行正規化處理，所得結果比較如下所示：

	Private Score	Public Score	RMSE
No Regularization	0.84964	0.85282	0.85123
Regularization	0.85603	0.86179	0.85891

雖然差異並沒有十分顯著，但正規化後所得的結果會較佳。

工工碩二 R05546030 **彭新翔**

5. 請討論你認為哪個 attribute 對結果影響最大？

若要將所有可能（總共有 106 個特徵，其組合數為 $\sum_{i-1}^{106}\binom{106}{i} = 81129638414606681695789005144063$）皆納入考慮，顯然以有限的時間內並不能夠全部皆顯示出來，何況需要再加上判斷的時間。但這部分應可以根據 Generative Model 模型中，計算各個特徵的權重後，以其權重(weight)代表其重要性，若以此為判斷依據，顯然是以 fnlwet 的係數較高尤為明顯。但另一方面，亦有發現 captain_gain 的權重值也不小，結合一般資本社會思維，資本資本越高通常代表能賺取更多的錢，若以此為判斷依據將會和其他次高的權重有所關聯性（如再次高的 age 特徵，年齡越大通常社會經歷較高，擁有高收入是正常狀況）。綜合上述結果，我認為 captain_gain 對結果的影響應為最大。