Homework 2 Report - Income Prediction

學號：b04502031　系級：電機二　姓名：施力維

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

這次手刻的 generative model 與 logistic model 在 Kaggle 上的準確率如下表所示，其中 feature 都有經過 normalize，也有做第二題(1)(2)的 Preprocessing：

	Public	Private
Generative	0.8456	0.8420
Logistic	0.8576	0.8462

就這個結果看來，Logistic 的準確度大約比 Generative 1%左右，然而 Generative 也足以過 Simple Base Line，表現的也還算不差。Generative 表現較差的原因可能是因為它的實作方式是猜測資料會呈現 Gaussian Distribution 的樣子，然而實際上我們並無法得知資料的分佈是否平均，或是在取得管道上有不小心讓某些群體的佔有比例增加，這些因素都會使得 Predict 的結果不準。反之，Logistic model 的彈性較大，並沒有做這樣的假設，只要參數 tune 好就能夠有好的 performance。

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

這次的 best model 是使用 sklearn.LogisticRegression 的套件實作，為單純的 LogisticRegression model。

Feature 的部分(1)扣除 education_num，因為此部分跟 education 的資料是相互對應的，如下圖所示：

Educ_num	1	2	3	4	5	6	7	8	9	10
Education	Preschool	1st-4th	5th-6th	7th-8th	9th	10th	11th	12th	HS-grad	Some-college
Educ_num	11	12		13		14	15		16	
Education	Assoc-voc	Assoc-acdm		Bachelors		Masters	Prof-school		Doctorate	

(2) 並將”country_?”合併到”country_USA”，”workclass_?”合併到”workclass_private”，將缺失的資料補成比例較高的類別(兩者都有 95%UP)。

(3) ”fnlwgt”、”age”、”capital_gain”、”capital_loss”、”work_hours”加上二到六次的項。

Training 的部分使用 L1 的 regulization，λ設為 1，solver 則是使用「liblinear」。

Training data 的準確率為 0.8617，其中切出 10%的 Validation data，其準確率為 0.8606，上傳至 Kaggle 上的分數為(Public, Private)=(0.8624, 0.8603)。

3. (1%) 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。(有關 normalization 請參考：https://goo.gl/XBM3aE)

本次實作的 feature scaling 是一般的標準化，使得資料平均為 0，標準差為 1，在這邊分別對 logistic、generative、best 三種 model 進行測試，資料皆有做過第二題(1)(2)的 Preprocessing，測試結果如下表：

	Un-normalized		Normalized	
	Public	Private	Public	Private
Logistic	0.8523	0.8410	0.8576	0.8462
Generative	0.8456	0.8420	0.8456	0.8420
Best	0.8611	0.8523	0.8624	0.8603

可以發現對於 Generative 而言，有沒有 Normalize 並沒有差別，然而對於 Logistic 跟 Best 都有明顯的影響，這是因為 Generative model 是計算模型的機率，在轉換成 Gaussian Distribution 時就已經有把 data 的標準差、平均等因素考量進去，以作出適合的 model。而 Logistic 是透過 gradient decent 來實踐，不同維度之間若是 scale 差距太大，會嚴重影響到 gradient decent 的成效。

4. (1%) 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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

這邊使用 hw2_best 的 model 來進行測試，分別對不同的λ進行 training，做出來的結果如下表：

λ	0.001	0.01	0.1	1	10	100	1000
No regularization	0.8605						
L1	0.8606	0.8614	0.8613	0.8616	0.8593	0.8524	0.7910

根據這個實驗結果，regularization 對於準確度的影響並不會很大，有沒有加上的差別大約是 0.1%，只有少數影響，這可能是因為大部份 feature 都是 1 次，而且都只有 1 或 0 兩個數字，不太能夠 overfitting，因此加上 regularization 並不會影響太多。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

這邊使用 hw2_best 的 model 來進行測試，測試方式分別為拿掉其中一個 attribute，並看哪項使得準確率下降最多(education_num 皆已拿掉)。

all	age	fnlwgt	education	marital_status	occupation	relationship
0.8617	0.8598	0.8617	0.8542	0.8620	0.8562	0.8604
race	sex	capital_gain	capital_loss	hours_per_week	native_country	workclass
0.8617	0.8616	0.8589	0.8606	0.8612	0.8611	0.8586

從上表可以觀察到，attribute 影響最多前三名分別是「education」、「occupation」、「workclass」，跟直觀上的結果相似。