Homework 2 Report Problem Set

學號：r07921052　系級：電機所碩一　姓名: 劉兆鵬

1. (1%) 請簡單描述你實作之 logistic regression 以及 generative model 對此 task 的表現，並試著討論可能原因。

	Logistic Regression	Generative Model
Public Accuracy	0.82140	0.81960
Private Accuracy	0.82060	0.82260
Training Accuracy	0.82635	0.82045

在這次實驗中，logistic regression 在 training data 以及 public testing 的結果優於 generative model，但於 private testing 的結果卻是低於 generative model．推測由於 generative model 是屬於 close form 的方式，模型對於各個資料於各項類別的機率是基於訓練資料的分佈所給定的，所以當 testing data 的分佈機率與 training data 符合時，則能夠獲得較好的效果．而 logistic regression 則是透過 gradient descent 的方式調整模型的參數，使得在訓練過程能夠以 training data 來進行參數的修正，盡可能地調整使模型的預測結果能夠更符合 training data，而並不是依照訓練資料的機率分佈而給定固定的參數，對於預測資料變異較大時，則會有 overfitting 的問題產生．

2. (1%)請試著將 input feature 中的 gender, education, martial status 等改為 one-hot-encoding 進行 training process, 比較其模型準確率及其可能影響原因。

	One-hot-encoding	Raw data
Public Accuracy	0.82140	0.81460
Private Accuracy	0.82060	0.81520
Training Accuracy	0.82635	0.82525

在這次實驗中，使用 one-hot-encoding 會優於直接使用 raw data 的結果，我認為若直接使用 raw data 來進行訓練時，會對數值產生大小關係的問題，例如 gender 的 1 為男性 2 為女性，就會給予女性大於男性的問題產生．故為將此問題消去，則需使用 one-hot-encoding 的方式，只需給定模型是否擁有此一項特徵，而不給定此特徵數值大小，則能夠更良好的使用各個類別特徵所代表的性質．

3. 請試著討論哪些 input feature 的影響比較大 (實驗方法沒有特別限制，但請簡單闡述實驗方法)。

Add Feature

PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
0.81965	0.7966	0.78475	0.78615	0.79025	0.7847

Drop Feature

LIMIT_BAL	MARRIAGE	PAY_0	PAY_2	PAY_3	PAY_4
0.8198	0.8197	0.8008	0.8202	0.8205	0.8198
PAY_5	PAY_6	PAY_AMT1			
0.81975	0.81985	0.8202			

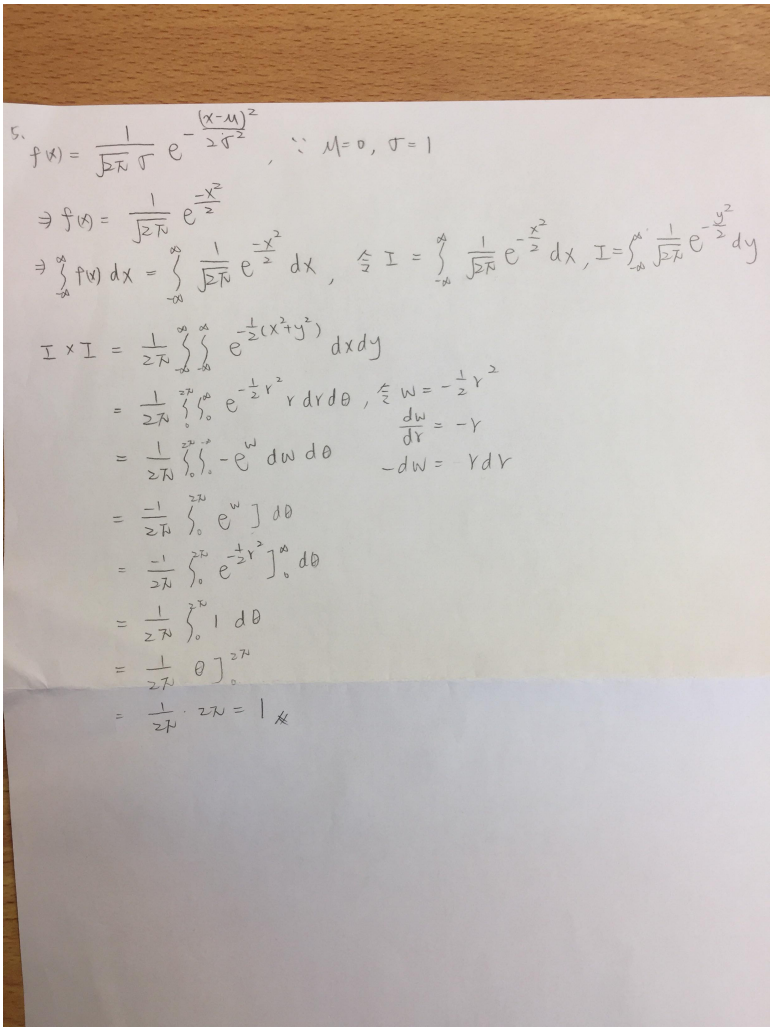
為了能夠了解各個 feature 的分佈對於此次 task 的影響程度，故我採用分別將各項 feature 逐一 Add 及 Drop 來進行訓練，並觀察其對於 validation data 結果，如上兩表所示．而在此會讓 validation data 的類別數目比與訓練資料的類別比相同，避免結果不一致．從逐一 Add 的 feature 中可以觀察到 PAY_0 至 PAY_6 為影響結果最大的 feature．而只以一項 Feature 訓練會使訓練參數過少，故又以逐一 Drop feature 來觀察其準確下降率，故可從表中得知 PAY_0 至 PAY_6、LIMIT_BAL、MARRIAGE 及 PAY_AMT1 最具有影響，將兩種方式結合起來就能察覺 PAY_0 至 PAY_6、LIMIT_BAL、MARRIAGE及 PAY_AMT1 的影響是比較大的．

4. 請實作特徵標準化 (Feature normalization)，並討論其對於模型準確率的影響與可能原因。

	Feature normalization	Raw data
Public Accuracy	0.82140	0.78040
Private Accuracy	0.82060	0.78120
Training Accuracy	0.82635	0.82355

沒有進行特徵標準化的結果是不理想的，認為這次作業的資料在類別的 feature 部分為 0 與 1 的分類值，而在其餘數值類的 feature 中，其數值為從零到幾萬的連續數值，若沒有使用標準化，則會讓模型訓練多偏重於數值較大的 feature，進而導致模型訓練不準確，故須先將各項數值經由 feature normalization 將值介於 -1 ~ 1 之間，並免模型訓練時發生不公平的評估結果．

Problem 5.



Problem 6.