ML HW2 學號：R05H41011 系級： 統計碩二 姓名：薛仲廷

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

Submission and Description	Private Score	Public Score
generative_prediction.csv 4 minutes ago by r05h41011_melo add submission details	0.84240	0.84520
logistic_prediction.csv 4 minutes ago by r05h41011_melo add submission details	0.85124	0.85393

由上圖看出 generative model 與 logistic regression 預測結果的 Pubic score 和 Private score 明顯都是 logistic model 較佳，其中 generative 的方法是將資料分為兩類並計算分別的 mean 和共同的 variance matrix，依照 Gaussian 分配去做 training；而 logistic model 有做 feature 正規化(min-max scaling)和用 adagrad 調整 learning rate。最後結果顯示為 logistic model 效果較好，我認為是因為這次的樣本足夠大所得出合理的結果。

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：
在 feature 上做不同的取法，將 feature 分成以下 6 類並組合起來訓練，其中設定 learning rate=0.7，iteration=10000

A=X_train[:,1:]、B=X_train[:,[0,1,3,4,5]]**2、C=X_train[:,[0,1,3,4,5]]**3、D=X_train[:,[0,1,3,4,5]]**4、E= np.log(X_train[:,[3]]+1e-100,axis=1)、

F= np.log(X_train[:,[0:3]]+1e-100,axis=1)

feature	*Training accuracy*	*Public score*
A+B+C	0.855233	0.85718
A+B+C+E	0.856390	0.85989
A+B+C+D+E	0.860561	0.86121
A+B+C+D+F	0.861582	0.86191

-

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。

答：

特徵標準化會讓每一個 feature 的大小相差變小，例如原始資料的前幾項和最後的目標變數 0、1 就差了許多，因此做特徵標準化會讓整個 training 效果更好。

	TRAINING ACCURACY	PUBLIC SCORE
MIN-MAX SCALING	0.85332	0.85393
NO FEATURE NORMALIZATION	0.78782	0.79841

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。

答：

Lamda	1000	100	10	0.1
Training acuracy	0.848741	0.849923	0.854834	0.854834

Regularization 對模型預測準確率沒有太大的影響，可能是因為模型不會太複雜導致 overfitting 的情況發生。

5.請討論你認為哪個 attribute 對結果影響最大？

答：

我覺得 capital 的影響最大，因為我模型中加入此 feature 後收斂速度快上許多，在模型準確度上也有明顯的提升。