# Machine Learning 2017 Fall
**Homework 02**

DUE DATE: OCTOBER 27, 2017 **學號：b03902129 系級：資工四 姓名：陳鵬宇**

**01. 請比較你實作的 Generative model、Logistic Regression 的準確率，何者較佳？**

Logistic Regression 較佳，在 Kaggle 上的結果如下：

1. Generative Model:
    - Private: 0.84240
    - Public: 0.84520
    - RMSE: 0.84380
2. Logistic Regression:
    - Private: 0.85628
    - Public: 0.86093
    - RMSE: 0.85860

**02. 請說明你實作的 Best Model，其訓練方式和準確率為何？**

我實作的 Model 為 Logistic Regression
訓練方式：adagrad
參數如下：learning rate = 0.05, iterations = 3000
其中有對 age, fnlwgt, capital_gain, capital_loss 以及 hours_per_week 做 2 到 5 次的 term append。
準確率 (Kaggle)：

- Private: 0.85628
- Public: 0.86093
- RMSE: 0.85860

**03. 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。**

在沒有對 feature 做 normalization 前，Kaggle 上的分數為：

- Private: 0.77066
- Public: 0.77125
- RMSE: 0.77096

在做完 normalization 後，Kaggle 上的分數為：

- Private: 0.85628
- Public: 0.86093
- RMSE: 0.85860

會有如此大的差異，或許是因為每一項的數值之間差異很大，例如 age 大約介於 18-65 歲
而 fnlwgt 可能是一個 5 位數的數字甚至是 6 位數
故若沒有做 feature normalization 的話，可能導致數字較大的那幾項 feature 影響結果較多。

**04. 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。**

在沒有做 regularization 前，Kaggle 上分數為:

- Private: 0.85100
- Public: 0.85454
- RMSE: 0.85277

在做完 regularization 後，Kaggle 上的分數為:

- Private: 0.85628
- Public: 0.86093
- RMSE: 0.85860

當做完 regularization 後 ($\lambda = 0.1$)，結果不論是在 Private 或是 Public 都變得更佳。

**05. 請討論你認為哪個 attribute 對結果影響最大?**

若要將所有 106 個 features 皆納入考量 ($\sum_{i-1}^{106} \binom{106}{i}$)，在有限的時間內是不太可能計算出來的，還需要再加上判斷之時間，所以我們可以運用 Generative Model 中，在計算各個 feature 的權重後，挑出權重較高的 feature，故我認為 fnlwet 的影響應較大，再來則是 age 及 capital_gain。