# Homework 2 Report - Income Prediction

學號：r06922095 系級：資工碩一 姓名：陳代穎

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

| | Private score | Public score |
|---|---|---|
| generative model | 0.84191 | 0.84557 |
| logistic regression | 0.84805 | 0.85712 |

不論是private或public score，logistic regression都有較好的準確率。有可能是因為在generative model的probability model我都使用Gaussian distribution，但事實上binary features應該使用Bernoulli distribution，因此增加了model的bias。

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

我利用support vector classifier (SVC)，我使用sklearn.svm下的SVC model。最終準確率為public score: 0.86412、private score: 0.85910。比logistic regression及generative model有較好的表現。

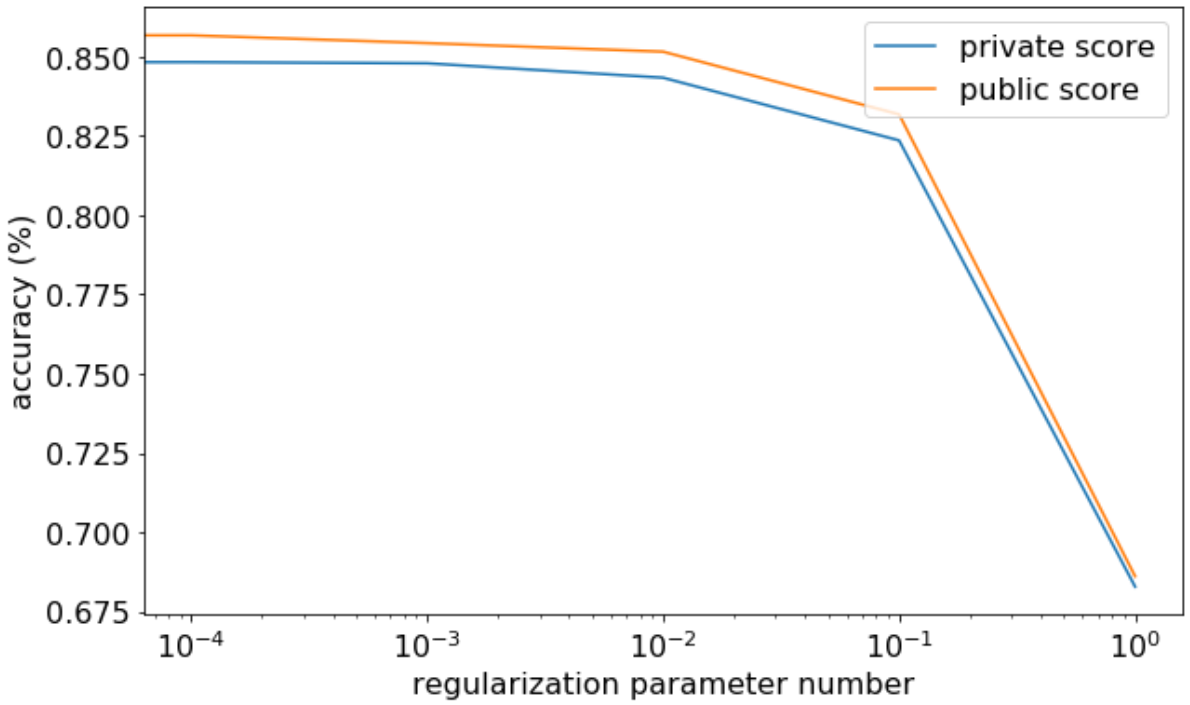
3. (1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。(有關normalization請參考：https://goo.gl/XBM3aE)

| | With normalization | | Without normalization | |
|---|---|---|---|---|
| | Private | Public | Private | Public |
| generative model | 0.84191 | 0.84557 | 0.84203 | 0.84557 |
| logistic regression | 0.84805 | 0.85712 | 0.77238 | 0.77788 |

normalization用standardization做使得data的mean為0、variance為1。可以看出normalization對logistic regression的影響大，因為logistic regression使用gradient descent的方式找解，因此每個維度的scale會影響gradient的大小，做過normalization後target function的圖形會更像一個圓，在做gradient descent時也比較容易走到local minimum。而generative model中standardization對probability distribution沒有影

響，因此generative model不會因為normalization而有所改變。

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。




分別用 $\lambda$ = 0.0001, 0.001, 0.01, 0.1, 1進行training，accuracy隨著 $\lambda$ 增加而減少，且private score與public score差異不大。因為regularization是用在overfitting的狀況，但因為linear logistic regression model並沒有overfitting的問題，若加上regularization只會使得training accuracy減少，造成testing accuracy的降低。$\lambda$ 越大regularization的影響就越大，從圖中便能看出此現象。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

| attribute | absolute correlation | correlation |
|---|---|---|
| education_num | 0.335154 | 0.335154 |
| relationship | 0.250918 | -0.250918 |
| age | 0.234037 | 0.234037 |

計算income跟所有attributes的correlation可以看出education_num這個attribute的correlation最大。而此欄位為受教育的時間，受教育時間越長則收入越高也是蠻合理的。