學號：B06901180 系級：電機三 姓名：鄭謹譯

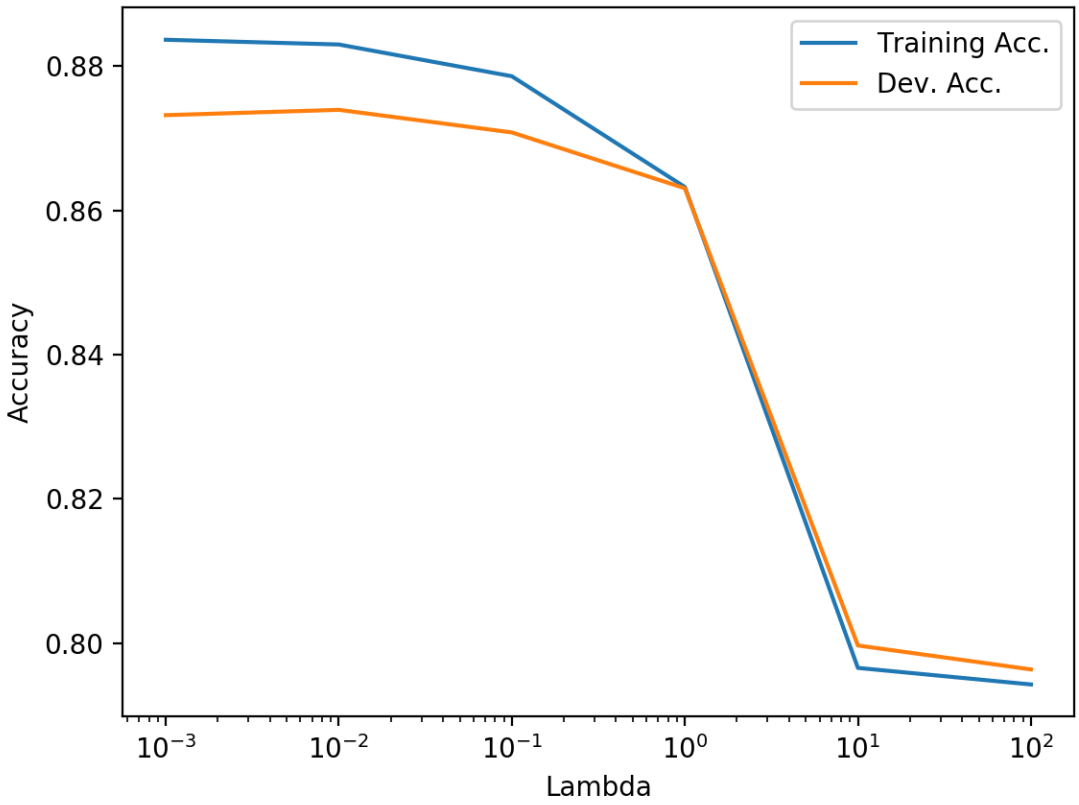
1. (2%) 請比較實作的 generative model 及 logistic regression 的準確率，何者較佳？請解釋為何有這種情況？

	Training Acc.
generative	0.872530
logistic	0.883617

Generative model 會自行建構出原本未給的資料與模型，將所有資料視為常態分佈，在資料量充足時此行為容易產生誤差 (underfit)，因此降低準確率。

2. (2%) 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。接著嘗試對正規項使用不同的權重 (lambda)，並討論其影響。(有官 regularization 請參考 https://goo.gl/SSWGhf p.35)

Lambda	0	0.001	0.01	0.1
Training Acc.	0.883617	0.883658	0.883002	0.878599
Dev. Acc.	0.873387	0.873203	0.873940	0.870807
Lambda	1	10	100	1000
Training Acc.	0.863240	0.796559	0.794266	Nan fail
Dev. Acc.	0.863068	0.799668	0.796351	Nan fail



由圖可知，training 和 dev. 的結果在 Lambda 大於 1 後皆驟降，可知過重的 penalize 會造成 training 過程誤差過大。但由圖與數據可發現，當 Lambda 介於 0.001-0.01 間時，準確度皆有略微提升，推測是因為 Lambda 讓 training curve 更平滑，使準確度些微上升，然而整體準確率並無太大差別。

3. (1%) **請說明你實作的 best model，其訓練方式和準確率為何？**

使用 logistic model 搭配 SGD 為基礎，經以下幾個步驟提升準確率。

(1) Feature Selection，大致計算相關性，最後刪除父母、自己國籍三欄資料。
(2) 針對數值項（非選題）增加 Feature，增加該項的 0.2 0.4 0.6 0.8 1.2 次方，小於 1 的次方能壓低特別大的數值，使大小中間的數值在標準化後比例放大，凸顯其價值。不同次方大小可以產生不同比例放大（1.2：縮小）的效果。
(3) Ensemble 對其初始W B 取 random，產生五組不同的 model，再使用

Logistic Regression 的方法 train 一個 model 利用五組 model 預測結果，然而效果沒有顯著提升，值得探討。

4. (1%) **請實作輸入特徵標準化 (feature normalization)，並比較是否應用此技巧，會對於你的模型有何影響。**

	Training Acc.	Dev. Acc.
Not normalize	0.762011	0.762440
Normalization	0.883617	0.873387

Feature normalization 減少 bias，同時除掉標準差可減少劇烈變化，以二（三）維來說讓 error space 接近圓（球），使機器更容易學出結果，結果有顯著的差別。