ML HW2 Report

學號 : B02901093 系級：電機四 姓名：吳岳

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

A:

先將train set做特徵標準化，其實就按照老師在Classification Model 裡面提到找P(C1IX)的方法代入，首先的目標是先將不同的答案(> 50K 薪水)分成兩群，個別求出106個維度的平均值跟Covariance Matrix。(如右圖)，

$$\mu^1 = \frac{1}{79}\sum_{n=1}^{79} x^n \quad \text{average} \qquad \Sigma^1 = \frac{1}{79}\sum_{n=1}^{79}(x^n-\mu^1)(x^n-\mu^1)^T$$

再找共同的Covariance。最後再將這些矩陣帶入Gaussian Distribution。即可運用z估測testing set的結果，最後training set的準確率為84.21%

$$z = \underbrace{(\mu^1-\mu^2)^T\Sigma^{-1}}_{w^T}x - \underbrace{\frac{1}{2}(\mu^1)^T\Sigma^{-1}\mu^1 + \frac{1}{2}(\mu^2)^T\Sigma^{-1}\mu^2 + \ln\frac{N_1}{N_2}}_{b}$$

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

A:

亦按照老師投影片的做法先算出sigmoid function f(x) = wX +b，weight隨機帶入， 再用右圖的式子不斷更新weight，直到lost開始增加或lost已趨近收斂(ex: lost變化量 < 10^-3)。

$$= \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right)x_i^n$$

Larger difference, larger update

$$w_i \leftarrow w_i - \eta\sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right)x_i^n$$

然而這樣訓練的成效真的有限，最好的狀況也才差不多 84.51%，無法更逼近Strong Baseline，於是我**再加上5個Contiuous Feature的三次方**，共111個feature(106+5)一起去訓練，準確率才突破85%大關。最後做的事情就是不斷微調regularization的lambda (0.05)跟rmsProp的alpha(0.7)，才總算突破Public Strong Baseline。

(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。

A:

以Generative來說，標準化後的準確率是84.21%，標準化前為84.24%。而當我將其中一些feature拿掉，做標準化的成效又反而比不做標準化還要好。例如說不考慮sex，標準化後為84.23%, 標準化前為83.20%。畢竟標準化最重要的目的是改善training時lost function遞減的成效，而Generative在運算當下就建好Model，不會牽涉訓練，故標準化改善成效有限，且也非正相關。

相對的，Logistic成效就比較顯著，在我還沒有加入3次方feature之前，未標準化的training 要上84%的準確率機率較低，而在標準化後，因有效降低decent的成本，得到84%準確率weight的機率較高。

(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。

A:

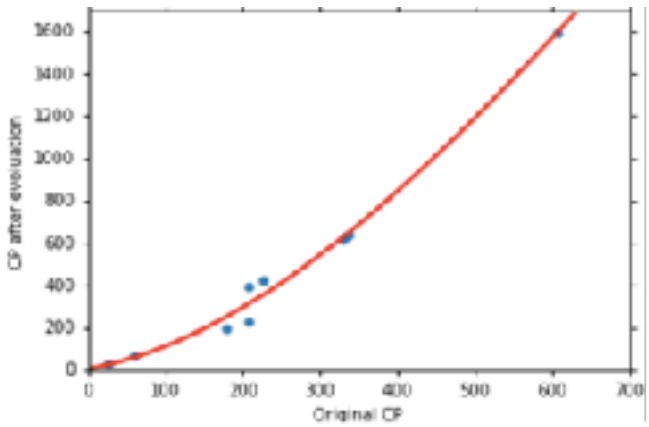
在training set狀況下，正規化讓準確率相對來說降低了(畢竟加入lambda)，而在某些lambda數值的條件下，testing set 的準確率會因weight的變化相對平緩而增加。

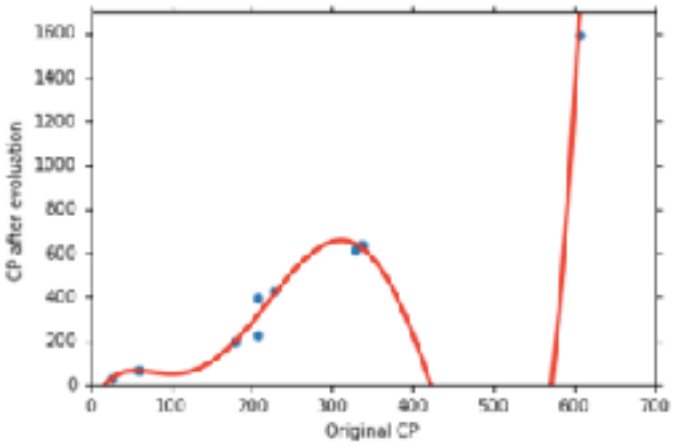
1. 加入三次方前

   原先設lambda = 0.5，但testing的準確率不盡理想，推測lambda落在右圖testing error開始增加的位置，而在調低lambda後testing set的最佳準確率提升了 0.5% (84.01% -> 84.51%)，lambda = 0.01。

2. 加入三次方後

   當初的直覺是次方提高lamda也要提高，畢竟高次方向的模型越容易藉由不合理的參數去overfit原本的training set，需要更高的lambda去讓高次方更平滑。(如下圖教授投影片三次方跟五次方的對照)，後來當lambda調到0.05時，也總算訓練出過strong baseline的模型。





(1%) 請討論你認為哪個attribute對結果影響最大？

A: Continuous Feature的三次方。

既然是連續變化，三次方的做法有效地凸顯了數據間的差異，讓準確率突破85%大關。其他的attribute(ex: rmsProp, regularization)對於準確率都有貢獻，但都比不上三次方直接貢獻的成效。