學號：R10921019　系級： 電機所碩一　姓名：鄭皓方

1. (1%) 請說明你是如何normalize discrete跟continous的feature

針對continuous的features，我利用下面的公式做normalize:

$$\frac{data - data.mean}{data.std}$$

至於離散的feature，我有對其中幾項進行處理:第一個是跟教育程度有關的變數，共11個。我把它們依據受教育的程度的高低整理成一個factor，如下:

Label name	Education degree factor
1st-4th, 5th-6th, 7th-8th, 9th	0
HS-grad"	1
Bachelors, Some-college	2
Assoc-acdm	3
Masters	4
Prof-school,Doctorate	5

另一個是根據婚姻關係安排的變數，我把它分成三個狀態，分別是有伴侶、曾經有伴侶，以及沒有伴侶。至於三個狀態在factor中的順序是選測試出來最好的。

Label name	Education degree factor
Married-civ-spouse, Married-spouse-absent, Married-AF-spouse	1
Widowed,Divorced, Separated, Widowed	0
Never-married	-1

針對剩下的feature，我利用 SKlearn 中的 select K best 函數挑選feature，這個函式是用univariate statistical tests找出對於output較顯著的特徵，主要是一些國籍和職業，以離散的方式放入先前整理好的資料進行訓練(共43個特徵)。

2. (1%) 使用DNN做 feature transformation，將output dimension設為4跟1024，並丟進linear SVM訓練，比較leaderboard上的結果，並說明造成這樣結果的原因 (hint: linear SVM本身是linear classifier，資料必須是linearly separable的資料)

Label name	kaggle score
output dimeansion 4	0.84597
output dimeansion1024	0.83587

經過訓練以及比對Kaggle上的成績之後，output 1024層的結果比4層的好。我覺得這個結過是合理的。SVM是一種線性分類器，維度越大，資料點間的距離越大，SVM越容易在中間找出邊界。舉例來說，如果在一維平面找不到一個邊界分隔資料，但是在二維平面卻有機會找到一條邊界(如case 1)。如果在二維資料找不到一個邊界，在一維也找不到(如case 2)。

因此適當的增加資料有助於分類器做分類。如果維度太高的話可能會train出一些沒有規律的特徵，即使懲罰值C_i不高，還是有機會變成noise影響結果，因此適當的選擇特徵維度也是很重要的。