Homework 1 Report - PM2.5 Prediction

學號：b05902008　系級：資工二　姓名：王行健

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

logistic :	public[0.85491]	private[0.84694]
generative :	public[0.84508]	private[0.84227]

比較generative model以及logistic regression可以發現無論在public或是private上，logistic regression的表現都較好。其原因在於logistic regression的sigmoid可以以標準化(將範圍限至於0~1)的權重省視每筆資料的loss，因此不易受單筆資料離分界線的遠近影響，因此也使得logistic regression能更公平的對待每筆資料。

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

DNRF :	public[0.87346]	private[0.86807]

hidden size = 1000
num epoch = 10000
num tree = 99
activation = relu
optimizer = Adam
batch size = 100
learning rate = 1e-3
best model是以deep neutron random forest訓練出來的。雖然說是deep，但由於收斂太慢，最後還是只做一層hidden layer。其原理簡單來講就是經過bagging使不同的neural network產生差異，最後再讓每個neural network做投票決定最後結果。

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。

normalised logistic :	public[0.84250]	private[0.84056]
normalised generative:	public[0.84508]	private[0.84215]

在做完normalization之後可以發現不論是logistic和generative的正確率都沒有顯著的影響，其原因可能在於大部分的資料採用one hot encoding，只有五個維度的數量具有意義，其中fnlwgt又被刪掉了。因此即使做normalization也無法有效控制整體資料的均衡性。對於logistic而言，另外一個可能的原因則是期中的sigmoid已經達到normalize的效果，因此再做normalize的效益不大。

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

lambda = 1e0 :	public[0.85343]	private[0.84535]
lambda = 1e2 :	public[0.85331]	private[0.84510]
lambda = 1e5 :	public[0.85368]	private[0.84682]
lambda = 1e10 :	public[0.83599]	private[0.83490]

可以發現當lambda<=1e5的時候，accuracy都沒有明顯的改變，但當lambda = 1e10，則出現underfit。從這兩點，不難判斷w的值其實相當小，因此即使不做regularlization，結果也不會太差。

5. (1%) 請討論你認為哪個attribute對結果影響最大?

整體而言，影響最大的應該是fnlwgt，在做logistic regression的時候如果不把這一項去掉，accuracy會落在0.76附近，而去掉之後，accuracy則大約0.85。再回去看fnlwgt，是代表這筆資料所能表現的樣本數量，直接用它來做regression並沒有意義。另外，實驗把fnlwgt當成權重也並不會得到更好的結果。