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

Model	Training	Valid	Test
generative (106維)	0.8432	0.8376	0.8431
logistic (106維)	**0.8543**	**0.8462**	**0.8523**
generative (114維)	0.854	0.8431	0.8527
logistic (114維)	**0.8593**	**0.8473**	**0.8568**

- 以上皆有將 training 和 testing data 一起進行過 z-normalization。
- 各資料的數量分別為 Train: 28061, Valid: 4500, Test: 16281 筆。
- Features分為直接使用助教提供的106維features，與加上幾個continuous features [‘age’, ‘fnlwgt', ‘capital_gain', ‘hours_per_week’] 的二次項與三次項變成 114 維兩種。
- 其中 generative model 套用 Gaussian Distribution 為機率模型，而 logistic regression 的相關參數為：learning rate = 0.1, batch size = 32, epoch = 250。未作 regularization。

可以看見 logistic 普遍比 generative 好一些，推測除了 logistic 可以做一些參數調整，透過 cross validation 盡量找出好 model 外，也可能是高斯分佈並不十分適合這次的 features，畢竟其中有101維都是 binary features，這件事也許也能從加上連續項的相關 features後 generative 就能通過 public & private 的 strong baseline 觀察到。

2.請說明你實作的best model，其訓練方式和準確率為何？

Model	Training	Valid	Test
gradient_boosting (114維)	0.9113	0.8683	0.8735

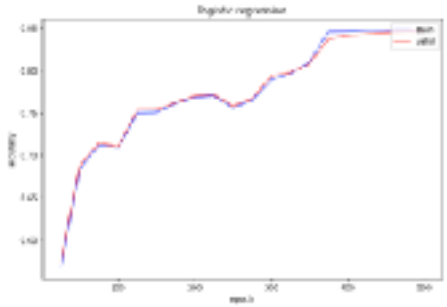
我的 best model 是利用 XGBoost 套件達成，其基於 GB (Gradient Boosting) 並進行許多優化，基本概念就是 iterative 生成多個 weak learner 然後綜合其結果。我使用的 booster 是 gbtree，較重要的相關參數包括：
'num_round': 1100, 'eta': 0.01, 'max_depth': 10, 'min_child_weight' : 1, 'max_features': 17, 'min_samples_split': 300
參數的調整方式是先固定較大的 eta 與小 num_round，依序對幾個參數做grid search 與 cross validation 最佳化。再將 eta 調小，cross validation 找出最好的 num_round，以求 model 表現得更穩定。

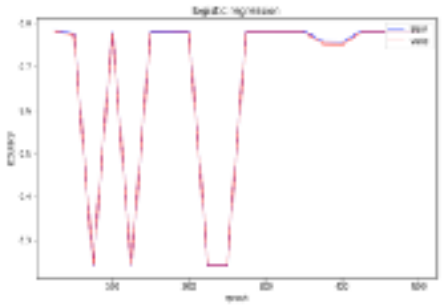
3.請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。

Model	Training	Valid	Test
generative (106維)	0.8434	0.8373	0.8432
logistic (106維)	0.8469	0.8456	0.846
generative (114維)	0.7644	0.7618	0.769
logistic (114維)	0.7813	0.7787	0.7843

以上皆為未作 normalization 的結果（有做的已在第 1 題），可以輕易發現不做會全面退步。尤其加上多次項後，部分維度可能過大 (如 capital_gain)，對模型造成強烈影響。另外，l_rate也需慎選，否則震盪將過於劇烈。

註：以上 logistic 為 500 個 epoch，若只做 250 在106 維會 underfitting，在114 維則會恰好遇到震盪的低點。可參考下圖（左為 106 維，右為 114 維）。





4. 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。

Model	Training	Valid	Test
λ = 0.1	0.8554	0.8473	0.8538
λ = 0.01	0.8586	0.8467	0.8553
λ = 0.001	0.8593	0.8471	0.8567

上述皆為對114維 features 的結果。可發現正規化對於模型的影響非常小，推測是此次模型本來就不複雜，不太有 overfitting 的問題。

5.請討論你認為哪個attribute對結果影響最大？

Feature	Test	Feature	Test	Feature	Test
age / fnlwgt	0.843/0.852	capital_loss	0.841/0.85	all_binary	0.78/0.80
sex	0.842/0.853	hours_per_week	X/0.851		
capital_gain	**0.836/0.838**	all_continuous	0.831/0.834		

上表是對 106 維抽掉各種 features 的一次項後的結果。可發現 capital_gain 影響最大。另外，抽掉 hours_per_week 會使 generative 無法計算 pseudo inverse。抽掉所有的連續或 binary 都會導致模型太簡單而使結果變差。