# 紀錄檢視



**2021/3/19 下午 05:59:12** 頁碼：21 撰寫者：王若宇(已查閱)

計畫代號：無

主題：First GAN's Data plot

關鍵字/標籤：python;ML;GAN

First GAN's Data plot

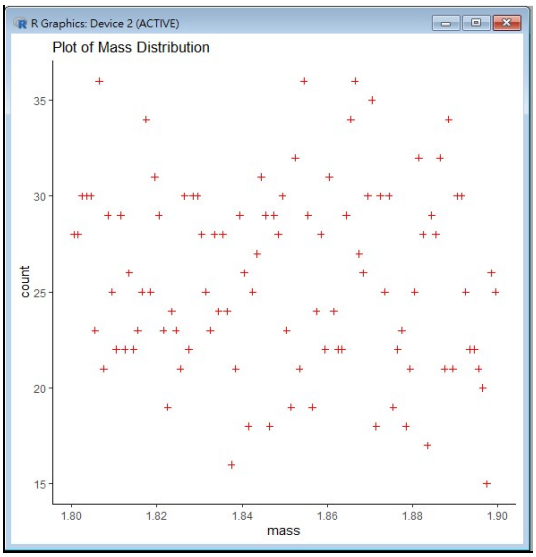
藉由gan生成的data現階段有兩個：tabgan兩種model，一個預設係數，另一個用上次的自訂係數，所以共有兩個資料。生成的時候。由於把entry當成是numerical feature來做訓練，entry會有超過原來存在的值，像是會有負數等等，因此會先去頭去尾，nPartical的部分，是來自於entry該類別的總數，因此訓練前先拿掉，待資料生成後，再數學計算添增回去。

而R 分析與繪圖的code，會因為features數量不匹配報錯，在gan生成的檔案再而外加上三行為0的features，以符合R code的import data shape。

以下為plot比較：

首先是原先的檔案的plot

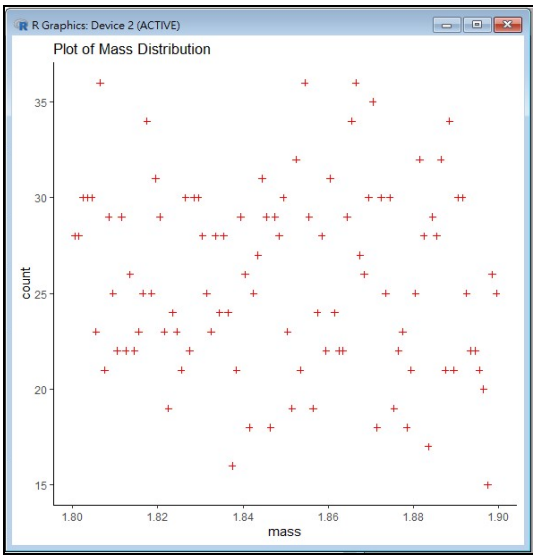
LEP1Data1992_recons_aftercut-MERGED.root.20200630.top100k




去除三個features再補添增三個為零的features回去

LEP1Data1992_recons_aftercut-MERGED.root.20200630.top100k_small_add3：

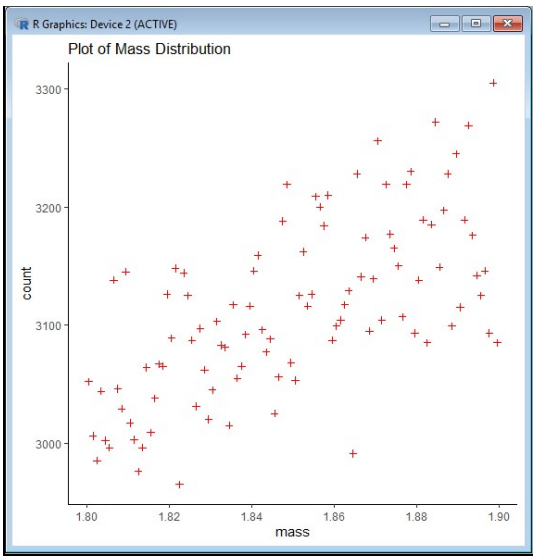
可以看到plot跟上一張圖一樣，plot的演算法與那三個features無關




以下是用3397entry資料做tabgan生成的資料

GAN_tabgan_sorted_cut_add3：

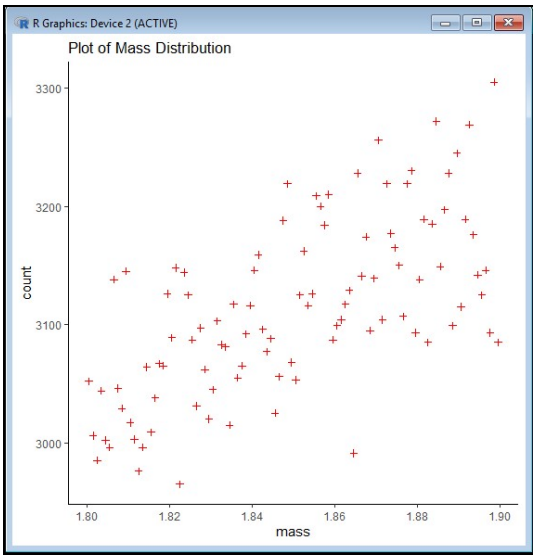
有1568429筆資料，沒有設定任何的模型係數，圖型有自己的分布，但是看起來跟原版有差距。




GAN_tabgan_detailed_sorted_cut_add3：

這組資料則是來自相同的package, tabgan，只是跟上面模型不一樣的點是有輸入其他的係數，但是plot總體看起來分布是差不多的

```
GANGenerator(gen_x_times=10, cat_cols=['pwflag', 'charge'],
             bot_filter_quantile=0.001,
             top_filter_quantile=0.999,
             is_post_process=True,
             adversaial_model_params={
                 "metrics": "AUC", "max_depth": 2,
                 "max_bin": 100, "n_estimators": 500,
                 "learning_rate": 0.02, "random_state": 42,
             }, pregeneration_frac=2,
             epochs=500).generate_data_pipe(train, target, test,
```




不管是訓練的原資料或是生成的資料，entry都是3397，之後會用10k entry的資料來做訓練，再來看看plot的分布。

全部展開 全部收合

參考文件(0)

查閱結果(1)

分享 (0)

挑選計畫編號

回到紀錄列表 回到查閱管理