

## 電子研究紀錄簿

系統時間(台北):2021/03/24 17:34:35 登入者:余承歡

管理與統計

撰寫入口

機密  
CONFIDENTIAL

## 紀錄檢視

首頁 &gt; 管理與統計 &gt; 查閱管理 &gt; 紀錄列表 &gt; 紀錄檢視

上一篇

比對相似度為0.00

檢視相似度比對結果

下一篇

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

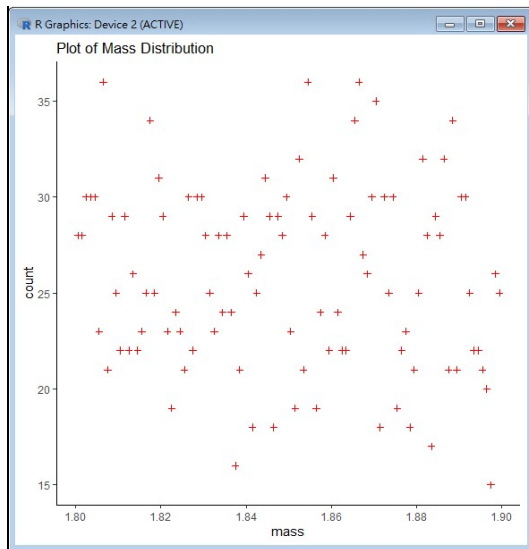
藉由gan生成的data現階段有兩個: tabgan兩種model, 一個預設係數, 另一個用上次的自訂係數, 所以共有兩個資料。生成的時候。由於把entry當成是numerical feature來做訓練, entry會有超過原來存在的值, 像是會有負數等等, 因此會先去頭去尾, nPartical的部分, 是來自於entry該類別的總數, 因此訓練前先拿掉, 待資料生成後, 再數學計算添增回去。

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

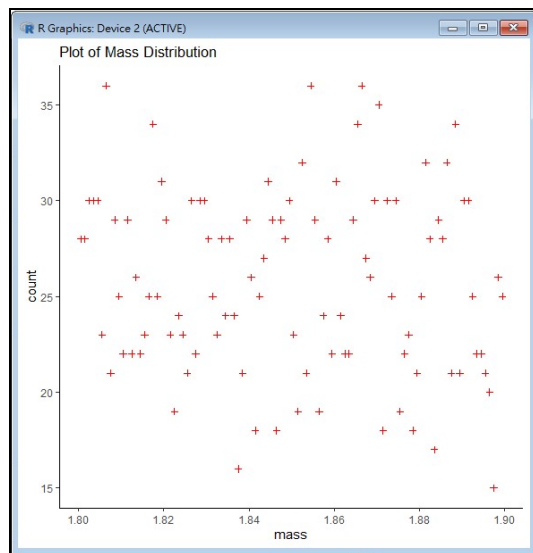
LEP1Data1992\_recons\_aftercut-MERGED.root.20200630.top100k



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:

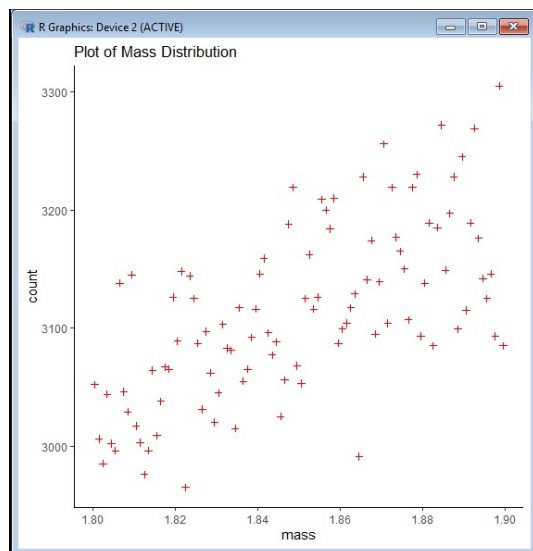
可以看到plot跟上一張圖一樣, plot的演算法與那三個features無關



以下是用3397entry資料做tabgan生成的資料

GAN\_tabgan\_sorted\_cut\_add3 :

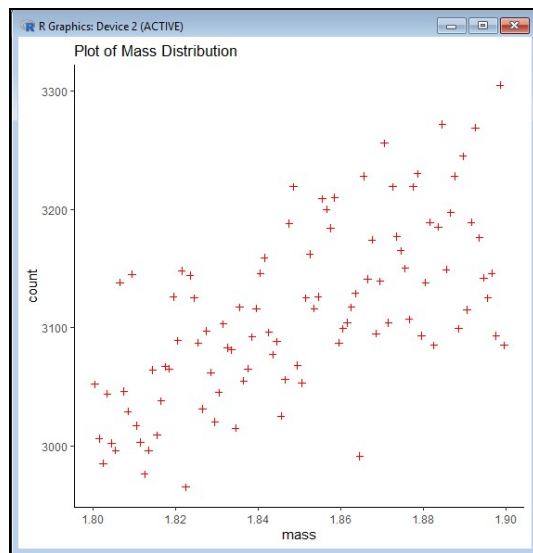
有1568429筆資料，沒有設定任何的模型係數，圖型有自己的分布，但是看起來跟原版有差距。



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,
             adversarial_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](#)

版權所有©2020 資訊工業策進會 | 限本會同仁公務使用禁止對外公開

財團法人資訊工業策進會機密資料 禁止複製、轉載、外流 III CONFIDENTIAL DOCUMENT DO NOT COPY OR DISTRIBUTE

業管部門 企推處 聯絡窗口：龔素嬌 02-6631-8614 資訊處 資訊服務櫃檯：電話：02-66318185 / [e-mail](#) / [Facebook](#) 系統使用說明書 | 撰寫入口