Data Mining Assignment 2

Author : Hzk

1. Data Creation

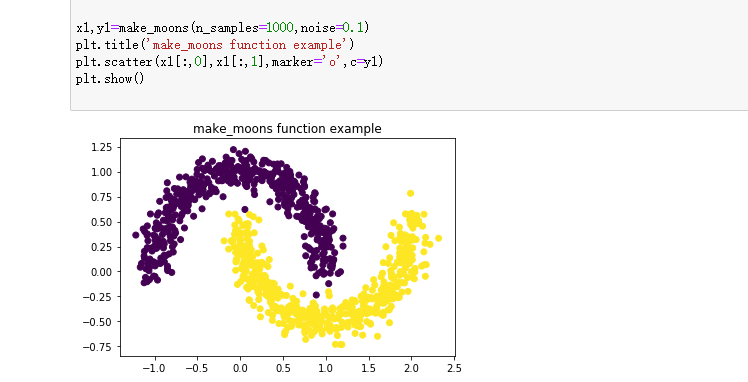
Report分為兩個部分，一部分是探索連續實數在線性可分性不斷變差的情況下各種分類器的表現。另一部分是決策樹在離散數據下的表現情況。

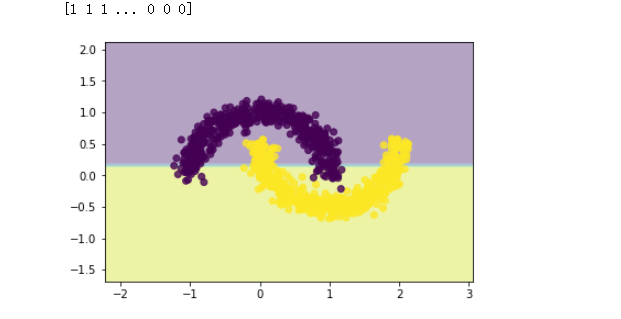
1. 離散數據在決策樹中的表現

Dataset1:

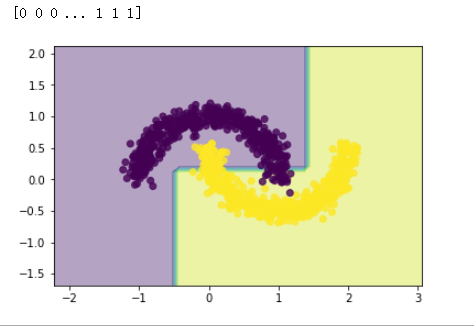
1. 連續實數在決策樹中的表現

Dataset1：月牙形數據，線性可分性較強

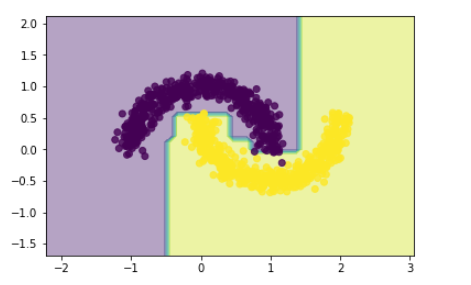
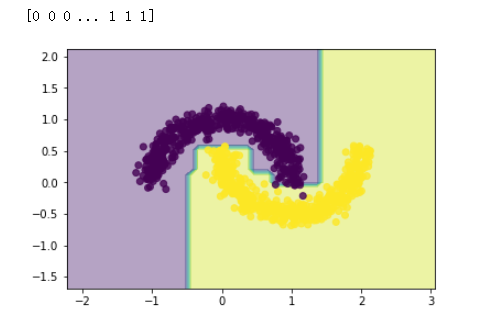


顯然在決策樹層數為1時，這就是一個線性分割問題，現在取決策樹層數為1，得到結果為

可以看到決策樹只是做了一次簡單的線性分割，現在把決策樹的層數加深，令層數為5，得到

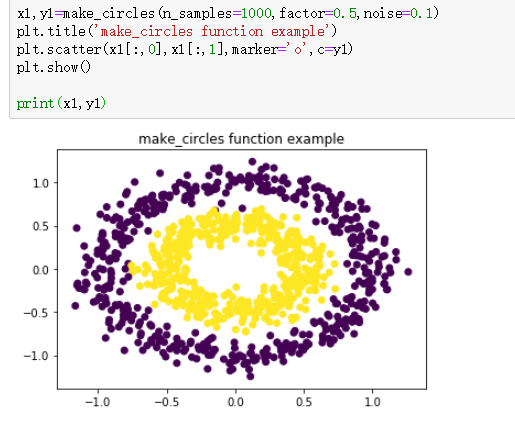


再取層數為10和20，得到

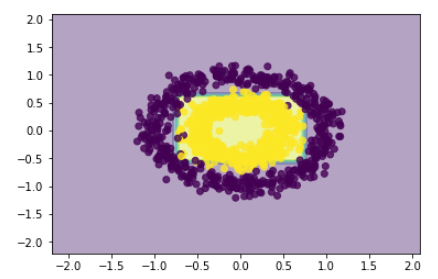
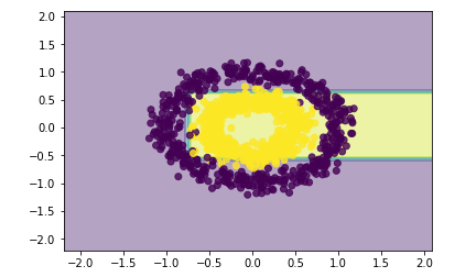
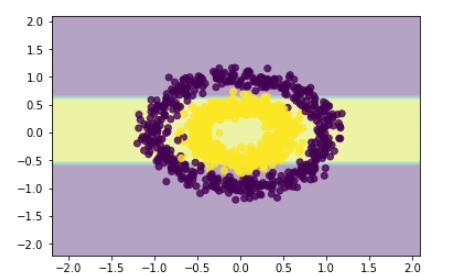


發現決策樹在10層已經收斂。

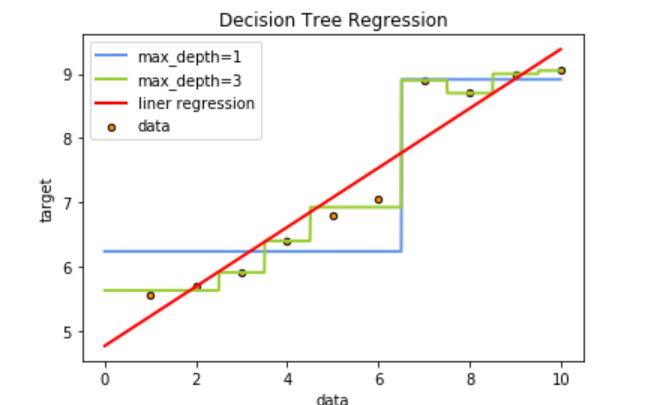
Dataset2：圓形包圍數據，線性可分析較差



下面三張圖是取決策樹深度分別為2,3,4時的分割情況，可以發現在二維空間中，二分類決策樹的margin就是多條直線相交組成的convex hull。



Dataset3：線性回歸模型



利用決策樹進行回歸，結果也相似。對於稀疏的線性數據，決策樹回歸的曲線也是一個有N個拐點的曲線，拐點個數隨著決策樹深度的增大而變多。