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

Dataset：蘑菇是否有毒

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 顏色 | 形狀是否奇怪 | 分泌物 | 是否有毒 |
| 1 | 鮮艷 | 是 | 無 | 有毒 |
| 2 | 不鮮艷 | 是 | 清亮 | 否 |
| 3 | 鮮艷 | 否 | 清亮 | 否 |
| 4 | 鮮艷 | 是 | 濃稠 | 有毒 |
| 5 | 不鮮艷 | 否 | 無 | 否 |
| 6 | 鮮艷 | 否 | 濃稠 | 否 |
| 7 | 鮮艷 | 否 | 濃稠 | 有毒 |
| 8 | 不鮮艷 | 否 | 無 | 否 |
| 9 | 不鮮艷 | 是 | 濃稠 | 有毒 |
| 10 | 鮮艷 | 是 | 清亮 | 有毒 |

將數據離散化可以得到新的dataset：

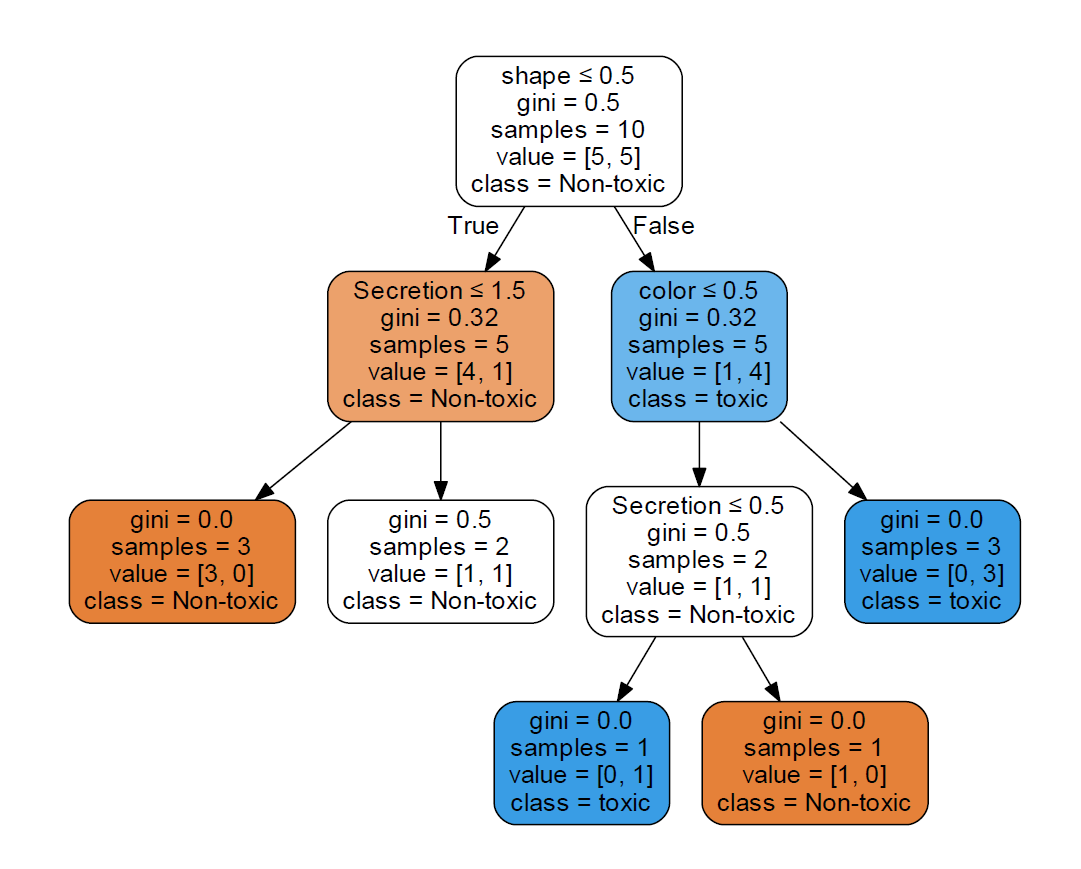
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 顏色（1=鮮艷，0=不鮮艷） | 形狀是否奇怪（1=是，0=不是） | 分泌物（0=無，1=清亮，2=濃稠） | 是否有毒（0=無毒，1=有毒） |
| 1 | 1 | 1 | 0 | 1 |
| 2 | 0 | 1 | 1 | 0 |
| 3 | 1 | 0 | 1 | 0 |
| 4 | 1 | 1 | 2 | 1 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 |
| 7 | 1 | 0 | 2 | 1 |
| 8 | 0 | 0 | 0 | 0 |
| 9 | 0 | 1 | 0 | 1 |
| 10 | 1 | 1 | 1 | 1 |

**數據產生的rule：**

1. 首先顏色鮮艷，形狀奇怪，分泌物濃稠中有至少2個符合的情況下，可以認為蘑菇是有毒的。
2. 在其餘case中，形狀奇怪的蘑菇，若含有分泌物（1或2），或者顏色鮮艷，也可以認為蘑菇是有毒。
3. 在形狀正常的蘑菇中，顏色鮮艷且含有分泌物的蘑菇也可能有毒。

使用這個Rule產生的結果中，所有結果都是正確的。

**決策樹決策過程**

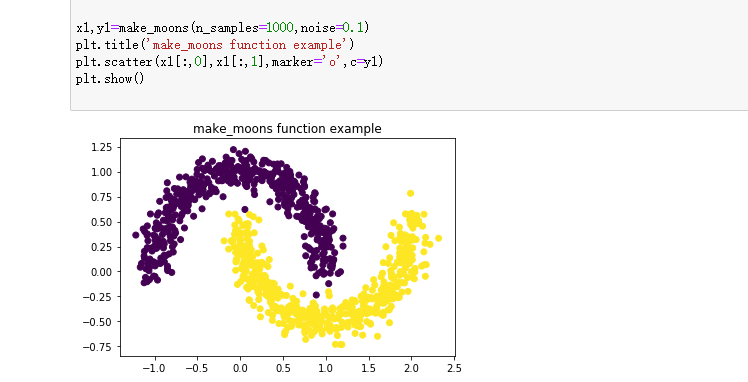


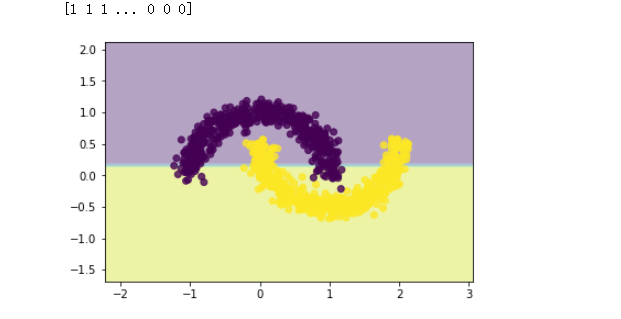
討論：

可以發現，決策樹也先通過形狀來進行第一層分層，在形狀奇怪的蘑菇中，會先選擇分泌物是否濃稠，在形狀不奇怪的蘑菇中，會先判斷顏色。使用這個決策方法得到的結果正確率也為100%。可以發現決策樹對能夠進行分類的變量反應還是比較靈敏的，在case不多且沒有noise的情況下，決策樹能夠很好的確定一個決策過程。

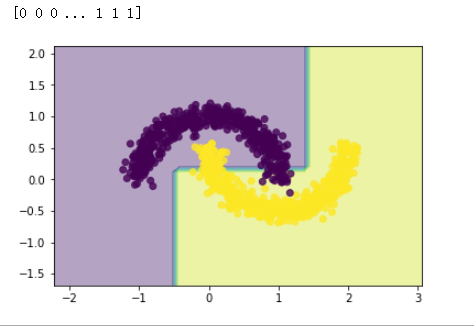
1. 連續實數在決策樹中的表現

Dataset1：月牙形數據，線性可分性較強

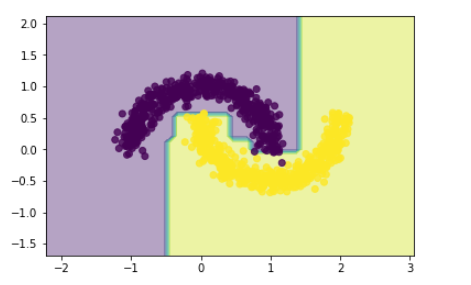
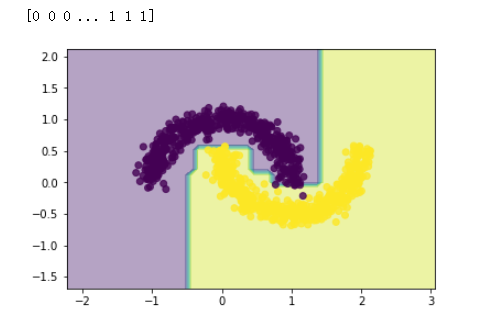


顯然在決策樹層數為1時，這就是一個線性分割問題，現在取決策樹層數為1，得到結果為

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

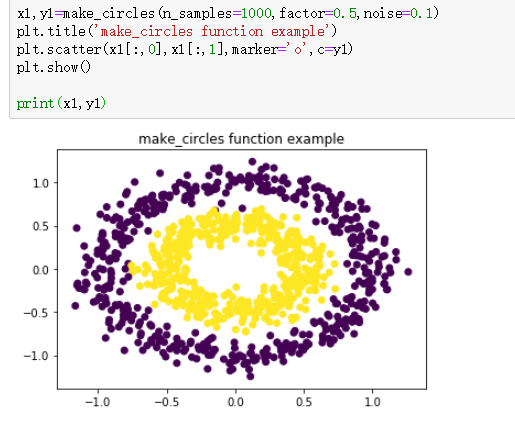


再取層數為10和20，得到

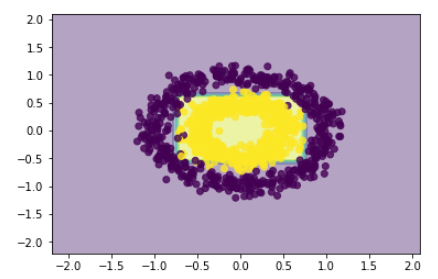
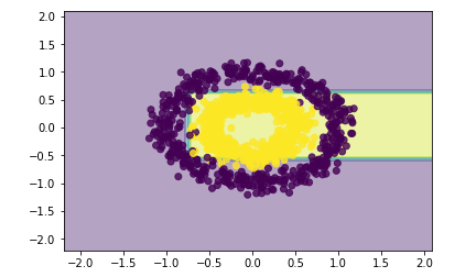
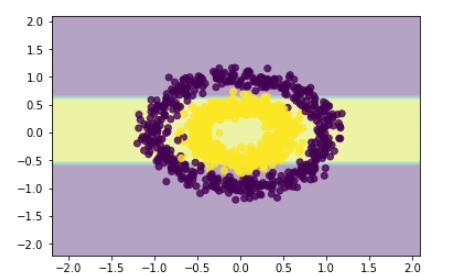


發現決策樹在10層已經收斂。

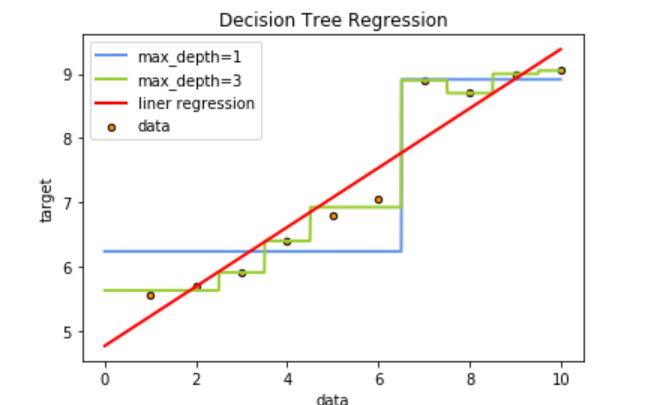
Dataset2：圓形包圍數據，線性可分析較差



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。