DATA MINING PROJECT TWO REPORT

2018/11/17

1. Absolutely right rule: Max\_length\_aspect\_ratio的value越大（170以上），Praxis\_aspect\_ratio的value越小（6以下），則越可能接近type 1（positive），否則為 -1（negative）.
2. 當樣本數量少但是feature多的時候，decision tree容易過擬合，一般來說，樣本數比特徵數多一些會比較容易建立strong的模型。當類別太多或者預設的最大深度越大時，運算時空量會明顯增加 。
3. 下圖是用pydotplus生成的決策樹視覺化圖，使用GINI指數來選擇最優劃分屬性，決策樹首選用到了資料X[5]、X[6]，即絕對Absolutely right rule裡的Max\_length\_aspect\_ratio和Praxis\_aspect\_ratio兩個features， 此時accuracy非常高。
4. 在加入了X[0]-X[4] 5個其他features之後，還是能看到決策樹用X[5]、X[6]還是作為優先特徵被用來分類。在max\_depth為7之後，model的accuracy隨之下降，但此時gini指數並未全部為0，證明X[0]-X[4]對最後的結果造成的影響並不是非常重要。總體來說，decision tree對於高緯度的feature也能進行比較好的處理。

