生科大四 C54046238 李承學

一、 環境

1. 作業系統
   Ubuntu 18.04
2. Python 版本
   Python 2.7.15rc1
3. 安裝的套件
   numpy-1.15.4 scikit-learn-0.20.0 scipy-1.1.0 graphiz-0.10.1
   backports.functools-lru-cache-1.5 cycler-0.10.0 kiwisolver-1.0.1 matplotlib-2.2.3
   pyparsing-2.3.0 python-dateutil-2.7.5 pytz-2018.7 setuptools-40.6.2 six-1.11.0
   subprocess32-3.5.3 python-tk2.7.15~rc1-1
4. 安裝的應用程式
   graphviz 2.40.1-2

二、 規則

利用溫度和濕度這二個 features，來預測上課是否會遲到。溫度和濕度都是介於 0-1 之間的一個連續值，會不會遲到有二種值，0 代表不會遲到，1 代表會遲到。總共有三種不同的規則，為了要讓這三種規則的 accuracy 可以比較，所以在設計規則時刻意設計使得會遲到與不會遲到的面積個佔一半，以下是這三種規則

1. 正方形
   （溫度 <0.5 且 濕度 <0.5）或（溫度 >0.5 且 濕度 >0.5）就會遲到，否則不會遲到
2. 等腰直角三角形
   （溫度 $+$ 濕度 $<0.5^{0.5}$）或（溫度 $-1+$ 濕度 $-1>-0.5^{0.5}$）就會遲到，否則不會遲到
3. 扇形
   （溫度$^2+$濕度$^2<1/\pi$）或（溫度 -1）$^2+$（濕度 -1）$^2<1/\pi$ 就會遲到，否則不會遲到

三、 檔案說明

rule_1、rule_2、rule_3 這三個路徑分別為三種不同規則的產生器以及所產生的資料，其中的資料檔名為 number_xxxx_noise_xx.csv，二個數字分別為樣本的數量以及雜訊所站的比例，用於計算 accuracy 的資料檔名為 test.csv。

主要的程式是 classifier.py，參數如下，其中若 CLASSIFIER 為 0 代表用 Decision tree，如果是 1 代表用 Random forest

```
usage: classifier.py [-h] [-c CLASSIFIER] [-t TEST] [-o OUTPUT] input

positional arguments:
  input                 the input csv file

optional arguments:
  -h, --help            show this help message and exit
  -c CLASSIFIER, --classifier CLASSIFIER
                                the classifier
  -t TEST, --test TEST  the csv file used to test accuracy
  -o OUTPUT, --output OUTPUT
                                the output png file of the rule of decision tree
```

下面的範例代表使用 Decision tree 來跑 rule_1/number_1000_noise_0.csv 這個檔案，並用 rule_1/test.csv 來計算 accuracy，將 decision boundary 繪製並顯示，decision tree 的規則可視化並且儲存至 tree.png

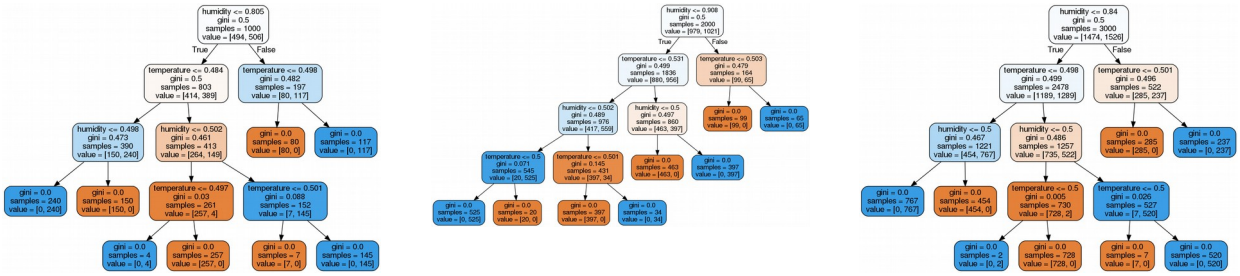
python classifier.py -c 0 -t rule_1/test.csv -o tree.png rule_1/number_1000_noise_0.csv

四、 結果

1. 利用 Decision tree 來跑 rule 1，並可視化。雖然 rule 1 的規則很簡單，如果是人類來分的話只需要二層就可以分出來了，但是在這邊卻分了很多層。雖然過程不一樣，但是結果卻很接近，accuracy 都很接近 1

Fig. 1 Decision tree rule 1 可視化

M = 1000 M = 2000 M = 3000



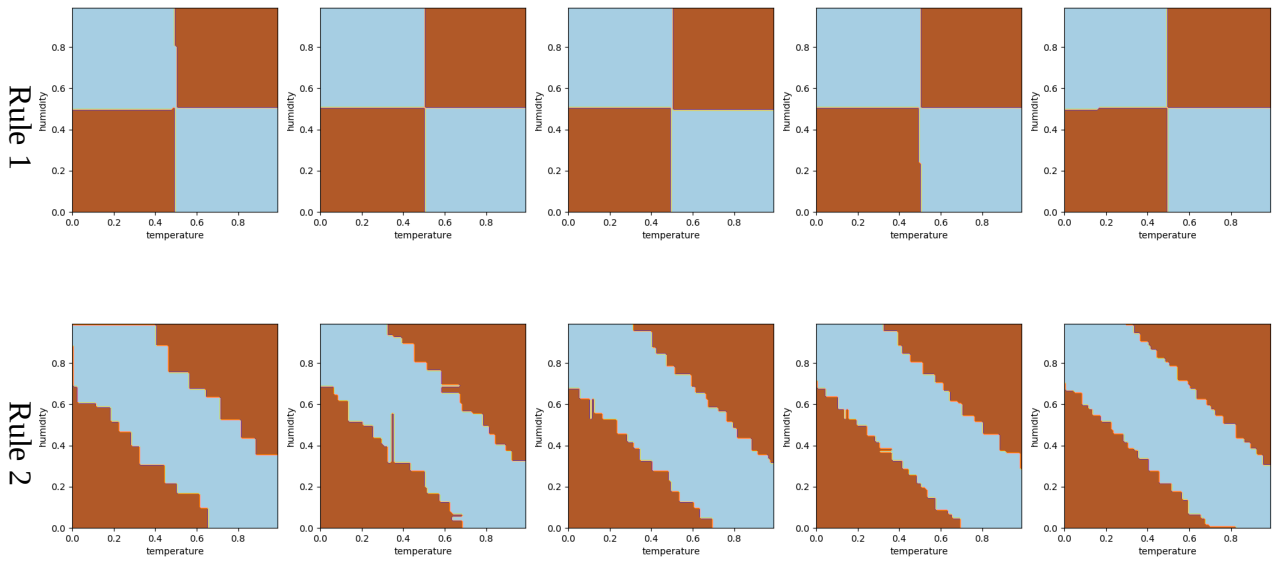
Accuracy = 0.99625 Accuracy = 0.99837 Accuracy = 0.99944

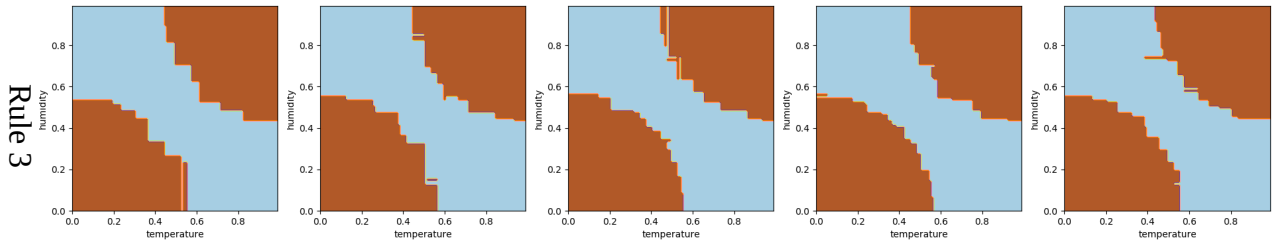
因為接下來的模型都很複雜，有很多層，所以就只畫 decision boundary

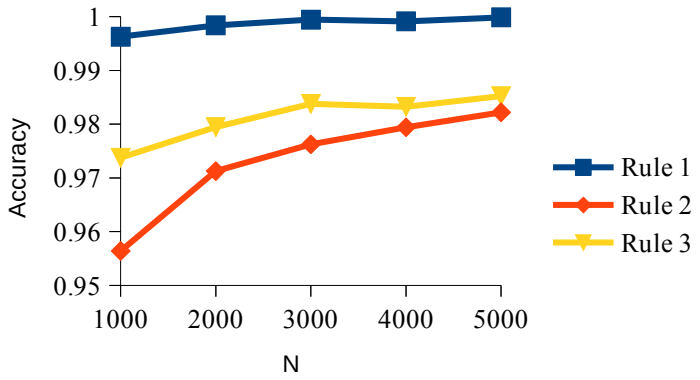
2. 測試不同規則，以及不同樣本數量下，accuracy 的差異。從 decision boundary 可以看出當樣本越多，邊界的鋸齒會越細微，而更接近一開始鎖定的規則。從圖中可以看出，對於 Decision tree 來說，rule 1 比較好訓練，在很小的樣本數下就有很高的 accuracy，其次是 rule 3，最後是 rule 2，猜測原因可能是 decision tree 只能用水平線及鉛直線來切割，rule 1 都只有水平和鉛直線，所以最好切割，rule 3 扇形的最旁邊也是接近水平和鉛直線，而 rule 2 完全就是 45 度的斜線，所以最難切割

Fig. 2 不同規則與數量下，所訓練出來的 decision tree

M = 1000 M = 2000 M = 3000 M = 4000 M = 5000



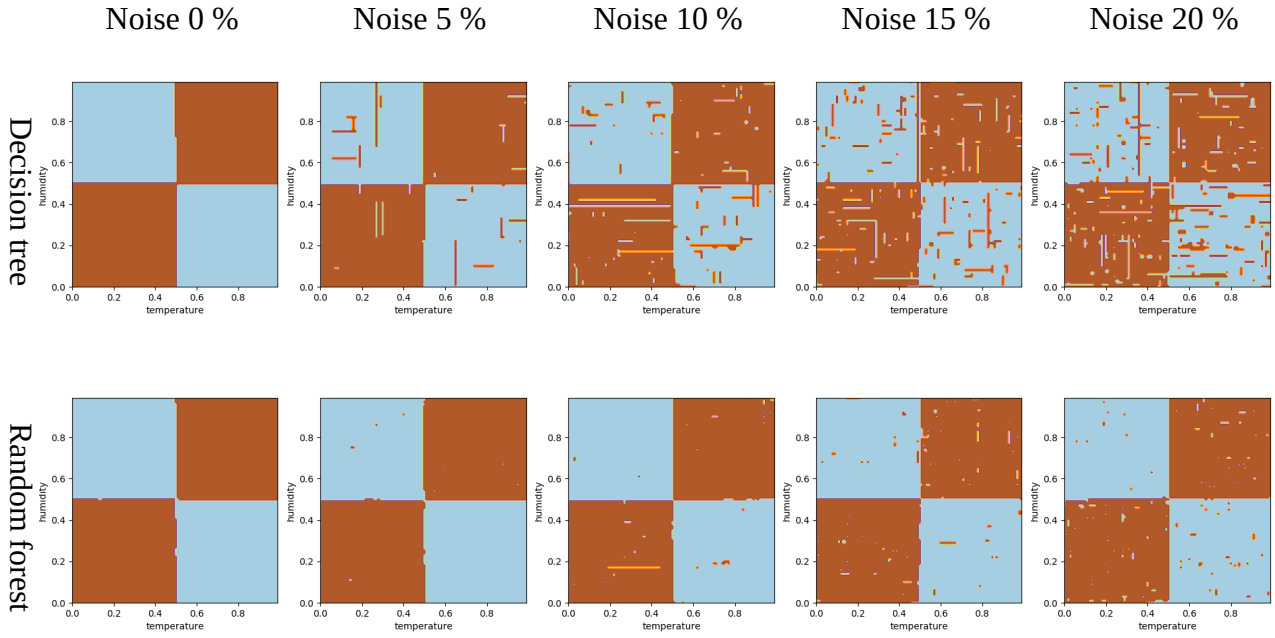


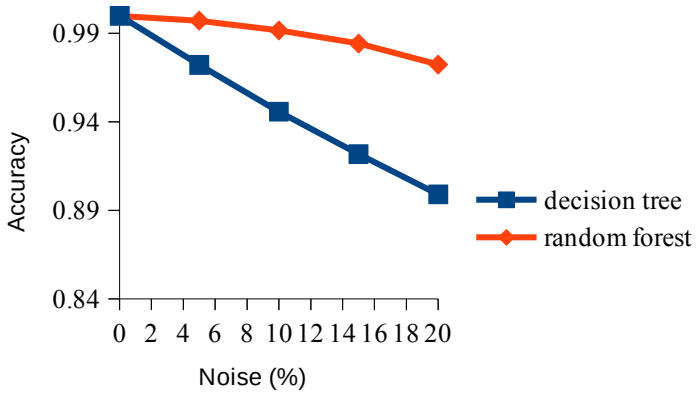


3. 對於以上的資料 decision tree 已經很夠用了，接下來要測試在不同程度的雜訊下，decision tree 的 accuracy 會有多少程度的降低，以及觀察 decision boundary 會有什麼樣的改變，並同時使用 random forest 與之比較。在雜訊逐漸增加的情況下，decision tree 會整個糊掉，而 random forest 則比較不會，原因是 random forest 是由很多個 decision tree 所組成，每個 decision tree 糊掉的地方都不太一樣，所以整體而言比較不會受到雜訊的影響。從 accuracy 也可以看出當雜訊增加時，dicision tree 的 accuracy 會下降較多，而 random forest 的 accuracy 則比較不會下降

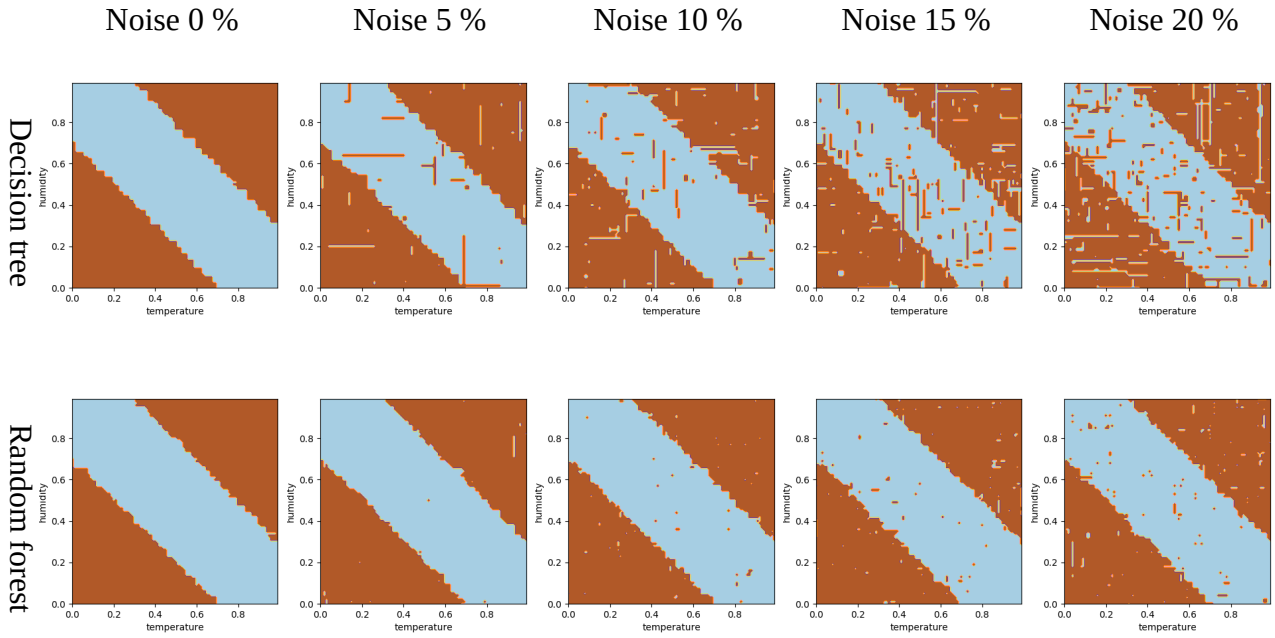
Fig. 3 對於每個不同的 rule，給予雜訊，觀察並且比較在不同雜訊下，decision tree 與 random forest 在 decision boundary 以及 accuracy 的變化

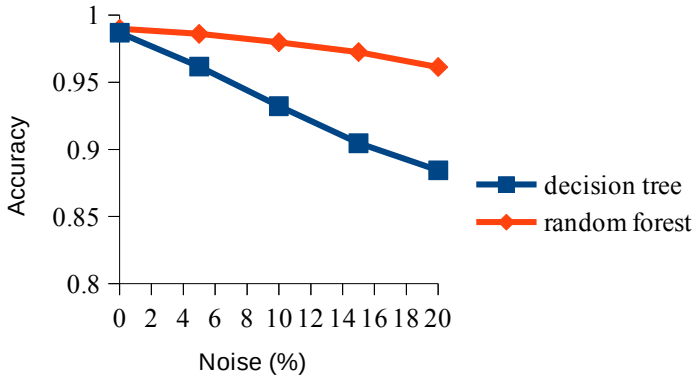
a. Rule 1



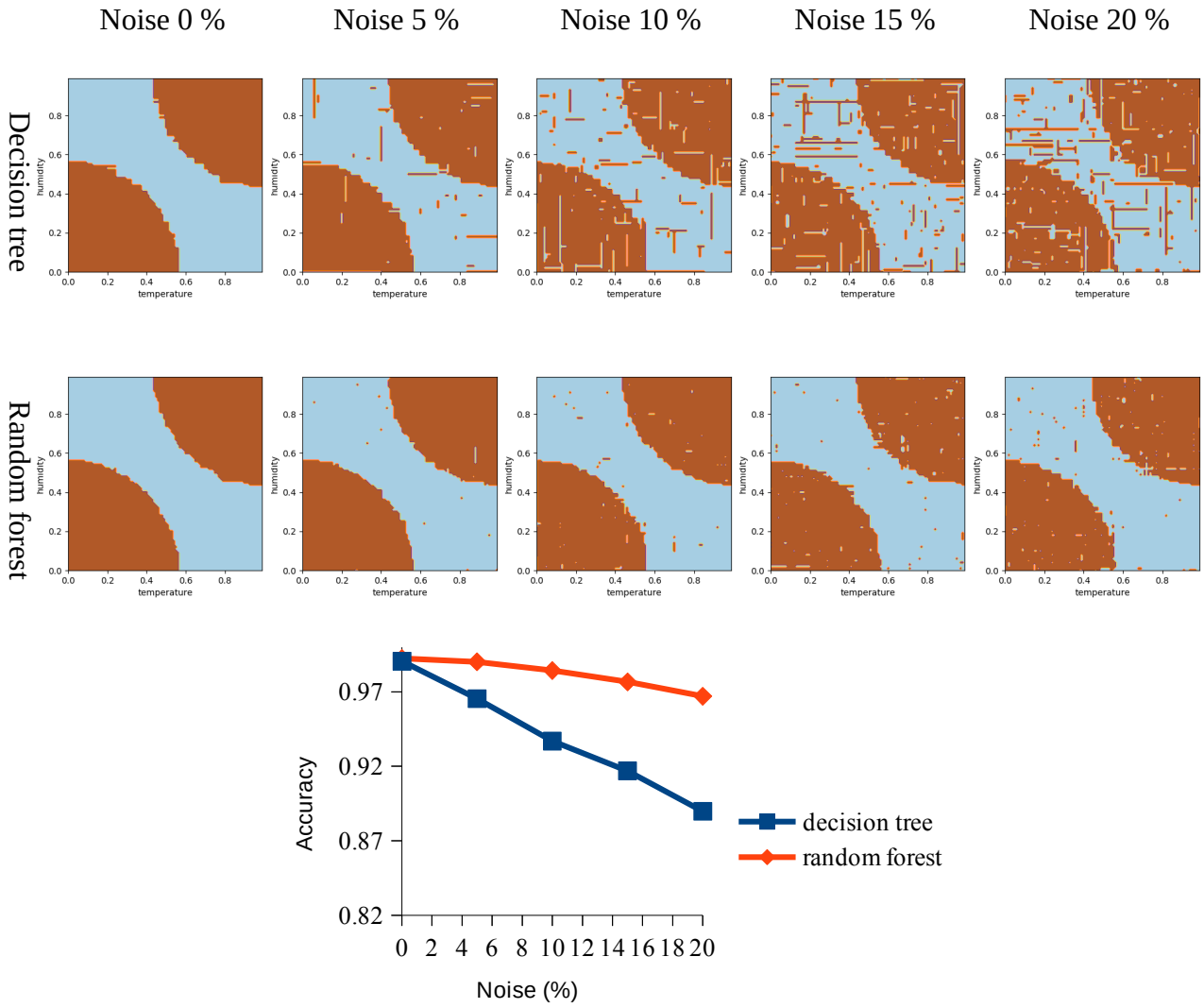

0.99
0.94
0.89
0.84
Accuracy
0 2 4 6 8 10 12 14 16 18 20
Noise (%)
decision tree
random forest

b. rule 2


Noise 0 %
Noise 5 %
Noise 10 %
Noise 15 %
Noise 20 %
Decision tree
Random forest
humidity
temperature


1
0.95
0.9
0.85
0.8
Accuracy
0 2 4 6 8 10 12 14 16 18 20
Noise (%)
decision tree
random forest

c. rule 3



五、 結論

1. 即使原本的規則很簡單，decision tree 有時後會用更複雜的規則來進行預測
2. decision tree 對於矩形的規則有較高的 accuracy
3. 當訓練資料越多，decision boundary 就越平滑，越接近原本的規則
4. decision tree 是一個容易解釋的模型，當訓練資料的邊界不是單純水平或鉛直線，會需要很多層判斷，雖然每一步都可以解釋，但是整體而言卻不容易理解
5. decision tree 很容易受到雜訊影響，但是 random forest 可以克服這個問題