國立成功大學

Data Mining 資料探勘

Project 2
Classification

姓名：金雅倫

學號：P96074105

目錄

一、目標說明 3
二、資料說明 5
三、內容說明 8
四、分析比較 9

一、目標說明<Goal Description >

概述

- **Goal**
  - Understand what classification systems do and the difference between real behavior of classification model and observed data
- Description
  - Construct a classification model to observe the difference between real 'right' data and modeled data

✧ 目標：

瞭解分類系統的作用，以及分類模型的實際行為與觀察到的資料之間的差異。

✧ 描述：

構建分類模型以觀察真實（絕對）"正確" 資料與建模資料之間的差異。

✧ 流程：

- **Step 1**: Design a set of rules to classify data, e.g., classify students with good performance.
  - You should design **k** features/attributes for your problems first.
  - Use 'absolutely right' rules to generate your positive and negative data (the number of data = **M**)
- **Step 2**: Use the data generated in Step 1 to construct your classification model
  - Decision tree is basic requirement, you can add more classification models.
- **Step 3**: Compare the rules in the decision tree from Step 2 and the rules you used to generate your 'right' data
- **Step 4**: Discuss anything you can

Step1：設計一資料集，需要 k 個特徵或屬性，使用“絕對正確”的規則來產生正負資料(數據數量=M)。

Step2：根據此資料集，產生新的分類模型。
(決策樹是基本要求，可以再依據個人需求新增額外更多的分類模型)

Step3：比較步驟 2 中決策樹所產生之規則和所自訂 "正確" 資料的規則之間的差異。

Step4：盡可能的針對結果討論任何想法。

二、資料說明<Data Description >

資料介紹

設計資料時，礙於憑空想像資料有些抽象，因此到政府資料開放平台(https://data.gov.tw/) 上尋找適合用來分析決策樹之資料。本次研究運用之資料為台南市 106 年度溺水地點水域之統計。戲水為國人夏秋之際一大放鬆休閒娛樂之活動，然而卻有很多因戲水意外溺斃的意外發生之事件。因此本研究預計運用此政府開放資料之相關資料屬性，例如：溺水的時間、地點、年齡、地標等稍作修改後訓練決策樹模型，以分析溺水之人員是否會被獲救。

✧ 資料來源：https://data.gov.tw/dataset/85838

原始之資料：

	A	B	C	D	E	F	G	H	I	J
1	年	月	日	時	溺水地點或附	水域種類	溺水結果	性別	年齡(歲)	
2	2017	1	5	20	五王大橋急流	溪河	死亡	男	34	
3	2017	1	7	16	玉港里西埔附	魚塭	死亡	男	78	
4	2017	1	30	16	秋茂園外海	外海(海岸緣	獲救	女	29	
5	2017	1	30	16	秋茂園外海	外海(海岸緣	獲救	女	26	
6	2017	1	30	16	秋茂園外海	外海(海岸緣	獲救	男	25	
7	2017	1	30	16	秋茂園外海	外海(海岸緣	獲救	女	24	
8	2017	1	30	16	秋茂園外海	外海(海岸緣	獲救	女	29	
9	2017	1	30	16	秋茂園外海	外海(海岸緣	獲救	男	54	
10	2017	2	2	20	金華新橋	溪河	獲救	女	43	
11	2017	2	3	21	四草大橋下游	近海(海岸緣	獲救	女	21	
12	2017	2	13	16	中華西路一段	溪河	獲救	女	50	
13	2017	2	13	18	天鵝湖	池塘	死亡	男	71	
14	2017	2	14	21	台84線東向西	魚塭	死亡	男	33	
15	2017	3	1	19	竹門里58-3號	圳溝	死亡	男	89	
16	2017	3	21	18	麻善大橋	溪河	死亡	男	35	
17	2017	3	26	6	環河街56號前	溪河	死亡	女	79	

✧ 轉換動作：將資料欄位修改並轉換為可分析之屬性。(將欄位之字串轉換為數字)

✧ 欄位說明：

欄位屬性	內容
time	時間：0 為晚上 / 1 為白天 17-24、01-07：晚上 / 8-16：白天
溺水地點	地點描述
water	溺水之水域或溪流名稱： 近海：1 /溪河：2 /碼頭：3 /外海：4 /圳溝：5
age	年齡：壯年：1 / 老幼：0 18-65 歲：壯年 / 其餘年齡：老幼
性別	男 / 女
survive	是否獲救 獲救：1 / 死亡：0

轉換後之資料：

A	B	C	D	E	F	G	H
time	溺水地點或附	water	age	性別	survive		
0	黃金海岸喜樹	1	0	男	0		
0	鹽水溪出海口	1	0	男	0		
0	蔡姑娘廟附近	2	0	男	0		
0	環河街56號前	2	0	女	0		
0	急水溪五王大	2	0	男	0		
0	台南市安平區	2	0	女	0		
0	安平區安平運	2	0	女	0		
0	營頂里新城橋	2	0	女	0		
0	安平商港21號	3	0	男	0		
1	汴頭里福如宮	5	0	男	1		
1	汴頭里福如宮	5	0	男	1		
1	嘉田里上茄苳	5	0	男	1		
0	竹門里58-3號	5	0	男	1		
0	176線9.5K(大	5	0	女	1		
1	三榮里嘉南大	5	0	男	1		
1	台南市安平區	1	0	男	1		

✧ 說明：

1. 所設計資料屬性欄位影響之因素：溺水時間(time)、溺水地點之水域或溪流名稱(water)、溺水人員之年齡(age)。
2. 不採納屬性原因：溺水之地點及溺水人員之性別，認為對於是否獲救並無太大直接影響，因此本次研究並不採納，但依然將欄位存放在資料裡。

✧ 設計之absolutely right rule：

觀察原始資料後發現：時間愈夜晚、光線較昏暗、溺水水域愈深、溺水人員年齡愈長者愈不容易獲救。因此以時間、水域、年齡為主要影響之屬性。

✧ 規則：

1. 時間是否為夜晚。
2. 年齡是否為老幼。
3. 水域是否為近海、溪河、碼頭。

三、內容說明<Description>

程式語言主要選擇使用 python，套件使用 sklearn 進行資料的前處理，首先將相關套件安裝並導入，再設定產生之 doc 資料檔案位置。

```
DM1110.py
1  import numpy as np
2  import pandas as pd
3  import os
4  from sklearn import tree
5  from sklearn import preprocessing
6  from IPython.display import Image
7
8  mypath = 'C:\\Users\\ellen\\Desktop'
9  os.chdir(mypath)
```

導入套件

使用 sklearn 套件處理

產生之資料位置

開始訓練決策樹，首先給定資料的屬性欄位，並設定樹的深度，最後設定訓練所要觀察的欄位以及輸出的資料數據檔案(doc)。→以本研究來說：要看是否會獲救，因此欄位為(survive)。

```
11 train = pd.read_csv("106.csv")
12 features = ["time","water","age"]
13 trainer=pd.DataFrame([train["time"],
14                       train["water"],
15                       train["age"]
16                         ]).T
17 tree_model = tree.DecisionTreeClassifier(max_depth = 3)
18
19 tree_model.fit(X = trainer,y = train["survive"])
20
21 tree_model.score(X = trainer, y = train["survive"])
22
23 with open ("tree3.dot", 'w') as f:
24     f = tree.export_graphviz(tree_model, feature_names=features , out_file = f)
25
```

訓練之 csv 檔

Data 欄位

樹之深度

訓練樹：欄位

輸出檔案

接著再把資料數據檔案 doc 檔之內容匯入，以產生決策樹之圖。

產生圖之線上資源：(http://www.webgraphviz.com/)

WebGraphviz is Graphviz in the Browser

Enter your graphviz data into the Text Area:

(Your Graphviz data is private and never harvested)

Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5

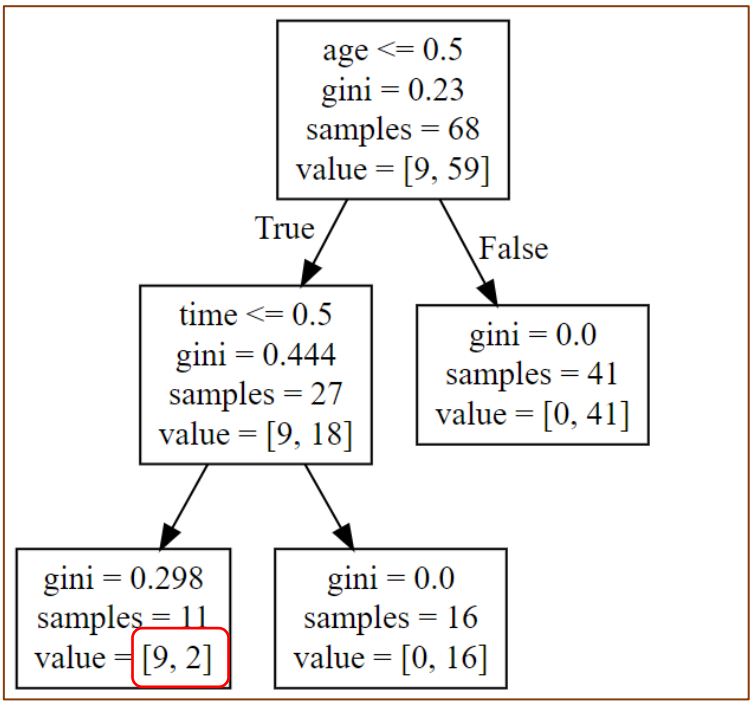
```
digraph Tree {
node [shape=box] ;
0 [label="age <= 0.5\ngini = 0.23\nsamples = 68\nvalue = [9, 59]"] ;
1 [label="time <= 0.5\ngini = 0.444\nsamples = 27\nvalue = [9, 18]"] ;
0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;
2 [label="gini = 0.298\nsamples = 11\nvalue = [9, 2]"] ;
1 -> 2 ;
3 [label="gini = 0.0\nsamples = 16\nvalue = [0, 16]"] ;
1 -> 3 ;
4 [label="gini = 0.0\nsamples = 41\nvalue = [0, 41]"] ;
0 -> 4 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;
}
```

深度為 2 之資料檔案

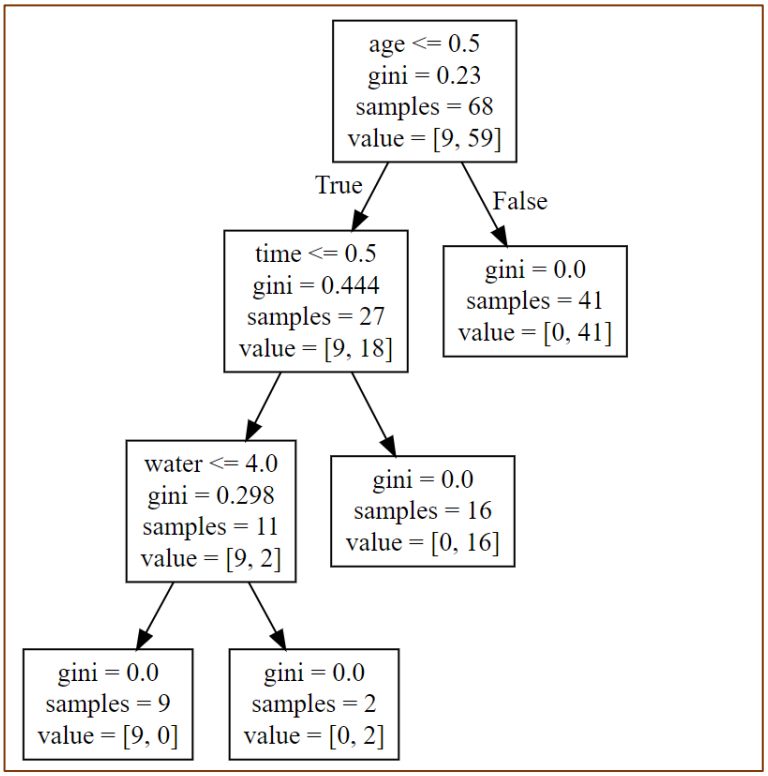
四、分析比較< Analyse & Compare >

此資料欄位屬性包括：時間、地點、水域、性別及年齡，分析時並無將所有的欄位都納入考量，因此，在分析時一開始將樹之深度設定 5 ，發現深度過於大，並無法看出有何異同，將數值一步一步往下修正後，可發現在 max_depth = 2 及 max_depth = 3 之分析結果存在些微差異。

✓ **深度為 2 之結果：**

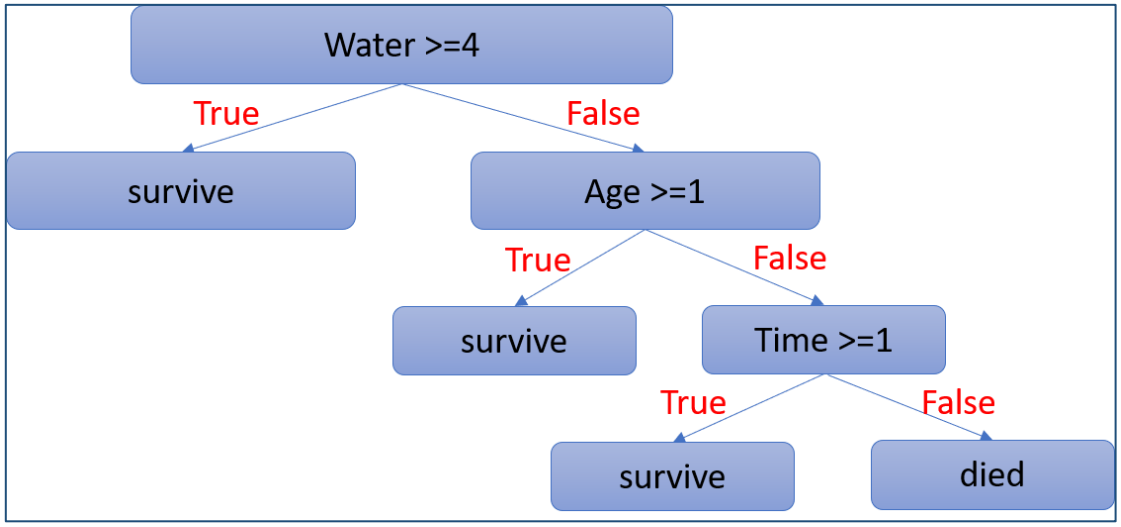


✓ **深度為 3 之結果：**



可以發現若設定決策樹之深度為 2，會造成決策樹還沒有分完的情形發生，因此只比較時間和年齡，並沒有比較到水域，出現還有 2 個屬於存活未分完之情形。而深度為 3 之決策樹，先比較年齡及時間，最後比較水域欄位是否為大於等於 4，挑選出最後有 9 筆資料為溺斃。

設計之正確規則中，決定是否存活之資料欄位為溺水水域(即 water 欄位)，因只要 water 數值大於等於 4 之資料，不管年齡或溺水時間都可以獲救(water = 4 外海、5 圳溝、6 魚塭)，預期決策樹如下圖所示。而決策樹則是先以年齡（age）為區分依據，只要 age 為大於等於 0.5，(age = 壯年：數值 1)之 ID，都會存活。因原始規則設定為交集合論，因此只要針對水域小於 4 之欄位中比較年齡與時間，若年齡為老幼、時間為夜晚兩者存在交集，就會溺斃。



發現決策樹訓練結果與原本所設計之 absolutely right 規則有些微落差，在設計之規則中，只要先比較水域是否為大於等於 4 之資料(即 water = 4、5、6)，再進一步比較時間、年齡屬性欄位，即可將存活數據判斷出來，而造成此一現象判斷原因為：所設計之規則不同所產生的先後判斷依據不同之現象，可以從程式碼訓練進行優化，給定主要決定結果之欄位值，依本研究為例，即為 water 屬性，因此，從水域屬性中往下進行條件決策，即可省去程式對資料進行重複篩選之過程。