Data Mining Project 2

P76061425 林聖軒

Environment

- DISTRIB_ID=Ubuntu
- DISTRIB_RELEASE=18.04
- DISTRIB_CODENAME=bionic
- DISTRIB_DESCRIPTION=Ubuntu 18.04.1 LTS

Usage

Classifier

```
$ python3 classifier.py [-h]
```

optional Options	Description
-h --help	show this help message and exit
-m METHOD	Classification method, dct=(Decision Tree),svm=(Support Vector Machine), default = dct
-train, TRAIN_PATH	Input training data file, default = ./data/train_data.txt
-test TEST_PATH	Input testing data file, default = ./data/test_data.txt
-k KERNEL	SVM kernel,default=rbf
-c PENALTY_C	SVM penalty parameter C of the error term,default=1
-cv CV	SVM cross_validate,default=10

- 用-m 來指定分類器，dct=(Decision Tree),svm=(Support Vector Machine)
- Decision Tree:

  用 Training Data 訓練 Decision Tree，並將訓練出的 decisionTree 結果 output 至當前目錄的 tree.pdf 中。
- SVM:

  用 Training Data 訓練 Support Vector Machine，並將 cross_validate 的結果及 Testing 的 Accuracy、Precision、Recall 輸出。
- 需要安裝 graphviz:

```
$ apt-get install graphviz
```

Data Generator

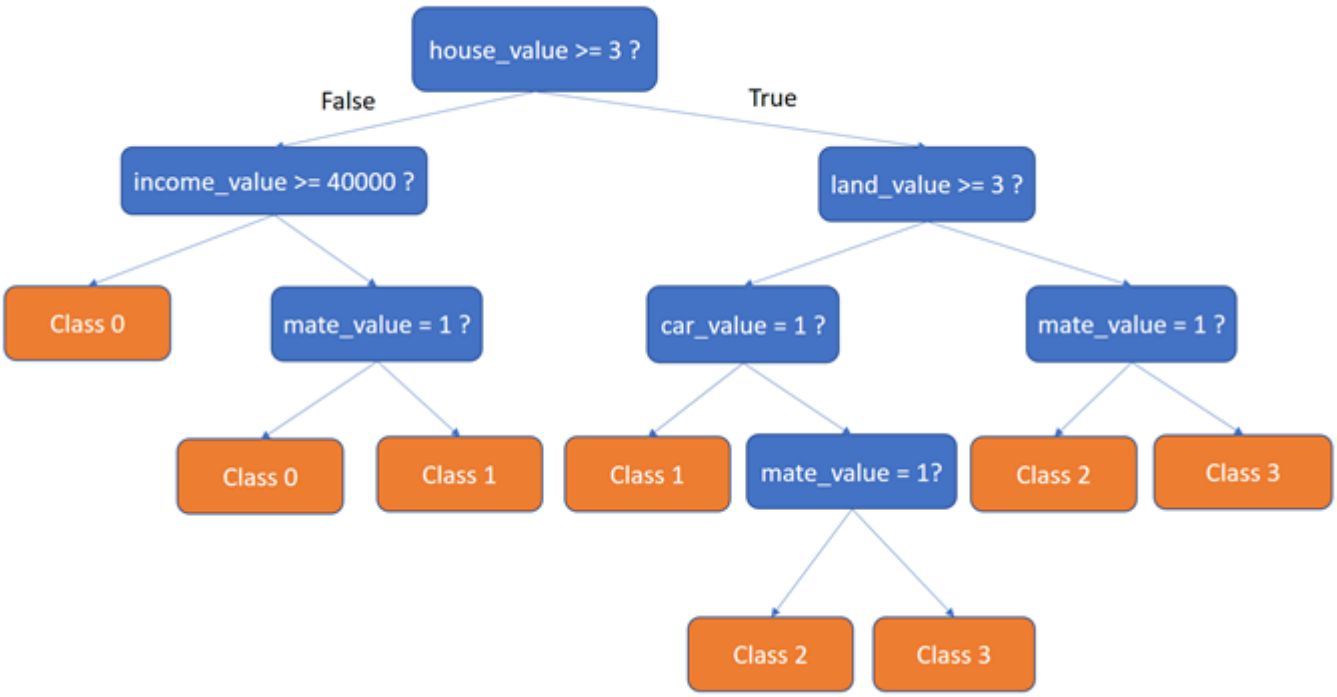
```
$ python3 data_generator.py [-h]
```

optional Options	Description
-h, --help	show this help message and exit
-train, TRAIN_SIZE	The number of training data you want to generate, default = 10000
–test, TEST_SIZE	The number of testing data you want to generate, default = 10000

- 執行後會在 data 資料夾內生成 10000 筆 training data(train_data.txt)及 testing data(test_data.txt)。

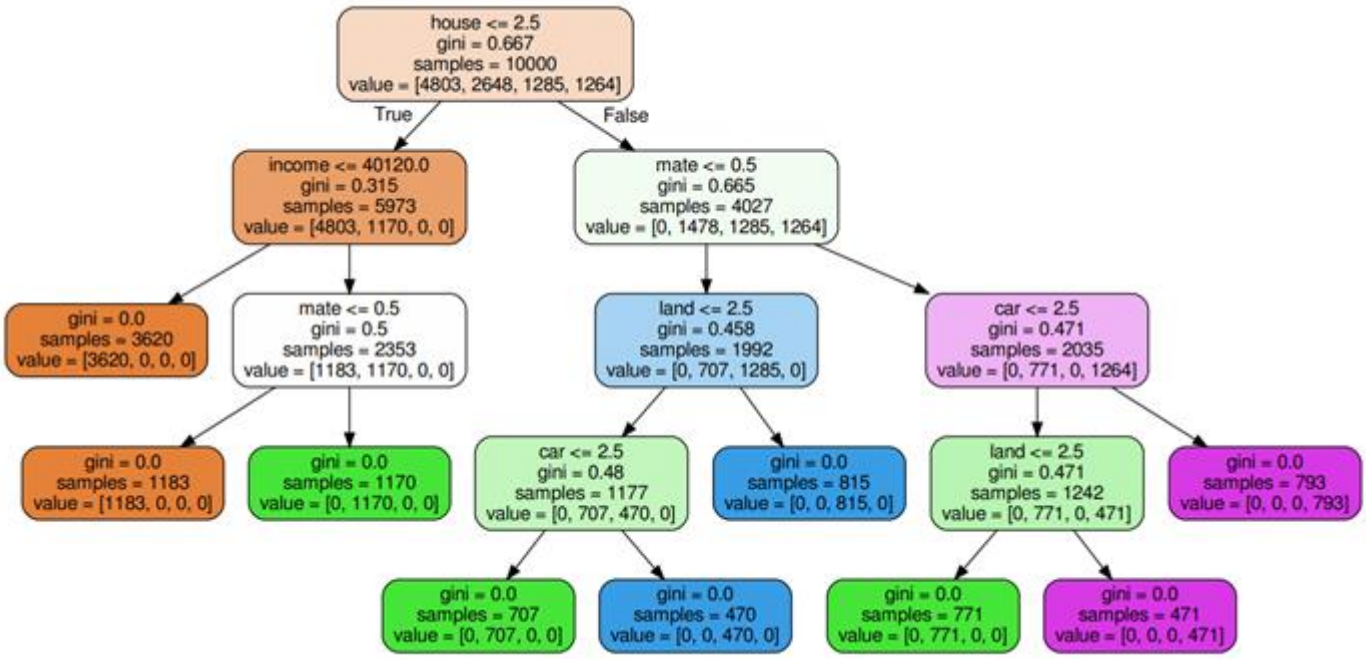
Absolutely Right Rules

- Attributes_list = house, car, land, income, mate, class
- house_value = [0, 5]
- car_value = [0, 5]
- land_value = [0, 5]
- income_value = [-50000, 100000]
- mate_value = [0, 1]
- class_vlaue = {0, 1, 2, 3}



Decision Tree

- Training size = 10000
- Testing size = 10000
- Criterion = Gini



Result metrics

- Accuracy = 0.9997
- Precision = 0.9999826689774697
- Recall = 0.9999677377726158
- 從結果圖可以得知，與 Absolutely Right Rules 相比，Decision tree 所建立出的 model 和實際的 rules 並非完全相同，但有很高的相似度，由於 Absolutely Right Rules 的規則很簡單，因此結果的 Accuracy, Precision, Recall 均有相當好的表現。

Support Vector Machine

- Svm kernel : rbf
- Penalty C : 1.0
- Cross Validate : 10

Cross Validation Result:

CV No.	Accuracy	Percision	Recall	Accuracy	Percision	Recall
Type	Train	Train	Train	Test	Test	Test
cv0	0.9971	0.9979	0.9979	0.9940	0.9961	0.9953
cv1	0.9970	0.9978	0.9979	0.9980	0.9982	0.9990
cv2	0.9974	0.9980	0.9983	0.9940	0.9949	0.9965
cv3	0.9969	0.9977	0.9978	0.9970	0.9976	0.9980
cv4	0.9966	0.9974	0.9976	0.9970	0.9984	0.9972
cv5	0.9961	0.9971	0.9973	0.9970	0.9976	0.9980
cv6	0.9968	0.9976	0.9977	0.9980	0.9990	0.9981

CV No.	Accuracy	Percision	Recall	Accuracy	Percision	Recall
cv7	0.9974	0.9981	0.9982	0.9940	0.9953	0.9960
cv8	0.9970	0.9977	0.9979	0.9950	0.9966	0.9962
cv9	0.9972	0.9978	0.9981	0.9970	0.9980	0.9976
avg	0.9970	0.9977	0.9979	0.9961	0.9972	0.9972

Testing Result:

- Accuracy: 0.9969
- Precision: 0.9976627899295014
- Recall: 0.9977905324777975
- 從結果可以看到，用 Support Vector Machine 進行分類的效果也非常好，但在簡單的規則下所定義出的 data，用 Decision Tree 這種簡單的 model 反而效果還要比用 SVM 來得更好，因此要視問題來決定 model，而不是一味的使用特定的 model 來解決問題。