School of Data and Computer Science, Sun Yat-sen University

E11 Decision Tree (C++/Python)

16337188 Ouyang Runyu

November 22, 2018

Contents

1 Datasets

The UCI dataset (`http://archive.ics.uci.edu/ml/index.php`) is the most widely used dataset for machine learning. If you are interested in other datasets in other areas, you can refer to `https://www.zhihu.com/question/63383992/answer/222718972`.

Today's experiment is conducted with the **Adult Data Set** which can be found in `http://archive.ics.uci.edu/ml/datasets/Adult`.

Data Set Characteristics:	Multivariate	Number of Instances:	48842	Area:	Social
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	14	Date Donated	1996-05-01
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	1305515

You can also find 3 related files in the current folder, `adult.name` is the description of **Adult Data Set**, `adult.data` is the training set, and `adult.test` is the testing set. There are 14 attributes in this dataset:

```
>50K, <=50K.


1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov,
State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm,
Assoc-voc, 9th, 7th-8th, 12th, Masters, 5. 1st-4th, 10th, Doctorate, 5th-6th,
Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated,
Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales,
Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct,
Adm-clerical,Farming-fishing,Transport-moving,Priv-house-serv,Protective-serv,
Armed-Forces.
8. relationship: Wife,Own-child,Husband,Not-in-family,Other-relative,Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
```

```
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany,
Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras,
Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France,
Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala,
Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong,
Holand-Netherlands.
```

Prediction task is to determine whether a person makes over 50K a year.

2 Decision Tree

2.1 ID3

ID3 (Iterative Dichotomiser 3) was developed in 1986 by Ross Quinlan. The algorithm creates a multiway tree, finding for each node (i.e. in a greedy manner) the categorical feature that will yield the largest information gain for categorical targets. Trees are grown to their maximum size and then a pruning step is usually applied to improve the ability of the tree to generalise to unseen data.

ID3 Algorithm:

1. Begins with the original set S as the root node.
2. Calculate the entropy of every attribute a of the data set S.
3. Partition the set S into subsets using the attribute for which the resulting entropy after splitting is minimized; or, equivalently, information gain is maximum.
4. Make a decision tree node containing that attribute.
5. Recur on subsets using remaining attributes.

Recursion on a subset may stop in one of these cases:

- every element in the subset belongs to the same class; in which case the node is turned into a leaf node and labelled with the class of the examples.
- there are no more attributes to be selected, but the examples still do not belong to the same class. In this case, the node is made a leaf node and labelled with the most common class of the examples in the subset.
- there are no examples in the subset, which happens when no example in the parent set was found to match a specific value of the selected attribute.

ID3 shortcomings:

- ID3 does not guarantee an optimal solution.
- ID3 can overfit the training data.
- ID3 is harder to use on continuous data.

Entropy:

Entropy $H(S)$ is a measure of the amount of uncertainty in the set S.

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

where

- S is the current dataset for which entropy is being calculated
- X is the set of classes in S
- $p(x)$ is the proportion of the number of elements in class x to the number of elements in set S.

Information gain:

Information gain $IG(A)$ is the measure of the difference in entropy from before to after the set S is split on an attribute A. In other words, how much uncertainty in S was reduced after splitting set S on attribute A.

$$IG(S, A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S \mid A)$$

where

- $H(S)$ is the entropy of set S
- T is the subsets created from splitting set S by attribute A such that $S = \cup_{t \in T} t$
- $p(t)$ is the proportion of the number of elements in t to the number of elements in set S
- $H(t)$ is the entropy of subset t.

2.2 C4.5 and CART

C4.5 is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. These accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rules precondition if the accuracy of the rule improves without it.

C5.0 is Quinlans latest version release under a proprietary license. It uses less memory and builds smaller rulesets than C4.5 while being more accurate.

CART (Classification and Regression Trees) is very similar to C4.5, but it differs in that it supports numerical target variables (regression) and does not compute rule sets. CART constructs binary trees using the feature and threshold that yield the largest information gain at each node.

3 Tasks

- Given the training dataset `adult.data` and the testing dataset `adult.test`, please accomplish the prediction task to determine whether a person makes over 50K a year in `adult.test` by using ID3 (or C4.5, CART) algorithm (C++ or Python), and compute the accuracy.
- Your codes should output the decision tree. The decision tree can be represented with a nested dictionary structure or be depicted by using the `plotTree` function in the book `Machine Learning in Action`.
- Hints (You can refer to the book `Machine Learning` written by Zhou):
  1. You can process the continuous data with **bi-partition** method.
  2. You can use prepruning or postpruning to avoid the overfitting problem.
  3. You can assign probability weights to solve the missing attributes (data) problem.
- Please finish the experimental report named `E11_YourNumber.pdf`, and send it to `ai_2018@foxmail.com`

4 Codes and Results

- You can see the code in Tree.py(with Chinese comment, coding=utf8)
- You can see the json of decision tree in Tree.json
- You can run the code with the following arguments:

```
parser.add_argument('--best_parameter', type = int, default = 0,
                    help = '若值为1，进行自动调参')
parser.add_argument('--post_pruning', type = int, default = 0,
                    help = '若值为1，进行后剪枝')
parser.add_argument('--print_tree', type = int, default = 0,
                    help = '若值为1，生成树的json文件')
parser.add_argument('--ignore', type = int, default = 0,
                    help = '若值为1，忽略第7、12、13个特征')
```

Figure 1: arguments

```
[nsccgz_yfdu_16@cpn21%tianhe2-K E11]$ python3 Tree_test.py
Read Data ...
Done ...
Train ...
Done ...
Test 10 times...

Predict 0 ...
Accuracy: 0.8507

Predict 1 ...
Accuracy: 0.8508

Predict 2 ...
Accuracy: 0.8508

Predict 3 ...
Accuracy: 0.8511

Predict 4 ...
Accuracy: 0.8514

Predict 5 ...
Accuracy: 0.8516

Predict 6 ...
Accuracy: 0.8509

Predict 7 ...
Accuracy: 0.8517

Predict 8 ...
Accuracy: 0.8508

Predict 9 ...
Accuracy: 0.8505

Total Accuracy: 0.8510288065843621
```

Figure 2: result(python Tree.py)

```
[nsccgz_yfdu_16@cpn21%tianhe2-K E11]$ python3 Tree_test.py --ignore 1
Read Data ...
Done ...
Train ...
Done ...
Test 10 times...

Predict 0 ...
Accuracy: 0.8563

Predict 1 ...
Accuracy: 0.8557

Predict 2 ...
Accuracy: 0.8565

Predict 3 ...
Accuracy: 0.8562

Predict 4 ...
Accuracy: 0.8563

Predict 5 ...
Accuracy: 0.8561

Predict 6 ...
Accuracy: 0.8563

Predict 7 ...
Accuracy: 0.8566

Predict 8 ...
Accuracy: 0.8559

Predict 9 ...
Accuracy: 0.8563

Total Accuracy: 0.8562189054726369
```

Figure 3: result(python Tree.py –ignore 1)

```
[nsccgz_yfdu_16@cpn22%tianhe2-K E11]$ python3 Tree.py --best_parameter 1 --ignore 1
Read Data ...
Done ...
Max Continous Son: 2, Max Leaf Length: 30
Train ...
Done ...
Predict ...
Accuracy: 0.8453

Max Continous Son: 2, Max Leaf Length: 31
Train ...
Done ...
Predict ...
Accuracy: 0.8462

Max Continous Son: 2, Max Leaf Length: 32
Train ...
Done ...
Predict ...
Accuracy: 0.8451

Max Continous Son: 2, Max Leaf Length: 33
Train ...
Done ...
Predict ...
Accuracy: 0.8467

Max Continous Son: 2, Max Leaf Length: 34
Train ...
Done ...
Predict ...
Accuracy: 0.8463

Max Continous Son: 2, Max Leaf Length: 35
Train ...
Done ...
Predict ...
Accuracy: 0.8477
```

Figure 4: result(python Tree.py –best_parameter 1 –ignore 1)

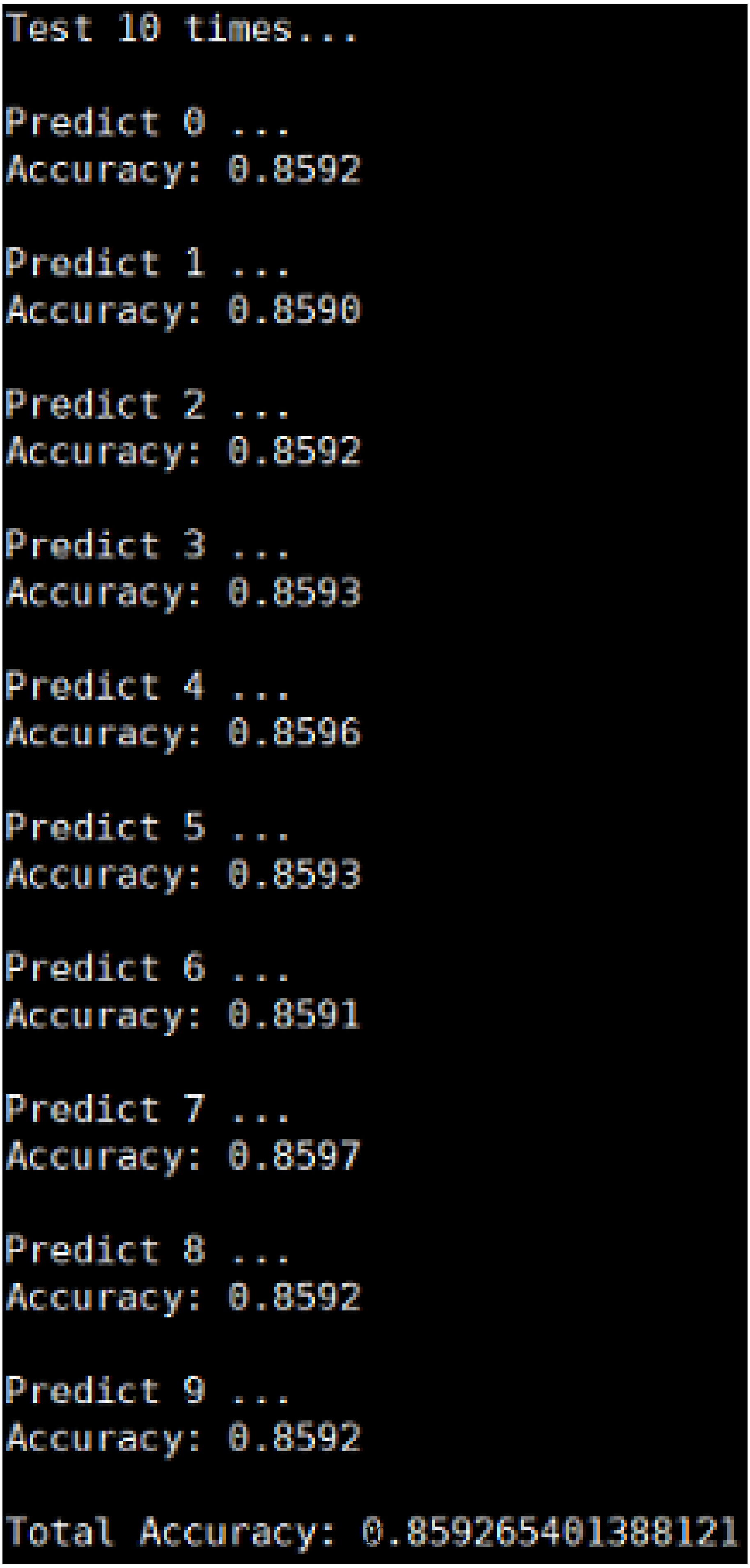

Figure 5: result(python Tree.py –post_pruning 1 –ignore 1)