I 526/B659 Programming Assignment 1 - Due Friday September 23, 2016

By

Mohit Galvankar - mgalvank@iu.edu

and

Abhishek Mehra -akmehra@iu.edu

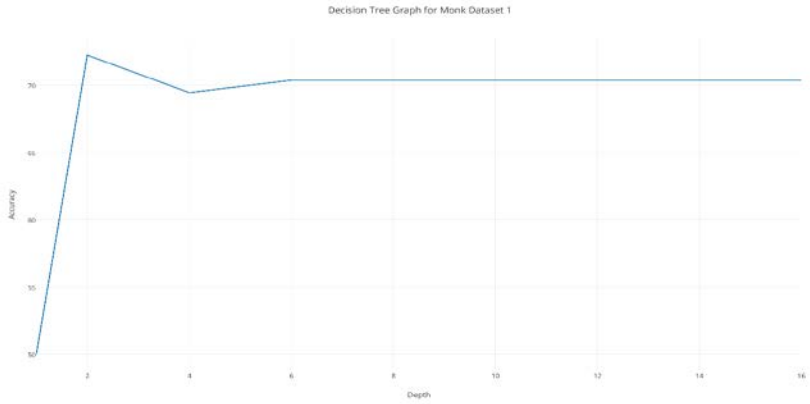
Implement a fixed depth decision tree algorithm. In particular, the input to your algorithm will include the training data set and the maximum depth of the tree. For example, if the depth is set to one, you will learn a decision tree with one test node, which is also called a decision stump. Test your implementation, with depth=1, and 2 respectively, on the following data set as described below (train on the training data and test with the testing data set). Data set information: This data set is extracted from the UCI Monk's problem data set (Monks-X.train and Monks-X.test). Note that there are 7 features (the first 7 columns) and the class labels are in the last col-umn. There are 2 classes. Please refer to http://archive.ics.uci.edu/ml/machine-learning-databases/ monks-problems/monks.names for details on the data set.

1.Start from depth = 1 and go to different depths (2,4,6,8...,16). For each depth, compute the error (the number of misclassifications) on the test set. Plot a learning curve with the depth of the tree on the x-axis and the accuracy on the y-axis.

- Monk 1 dataset

Depth	Error
1	0.5
2	0.27
4	0.30
6	0.29
8	0.29
10,12,14,16	0.29

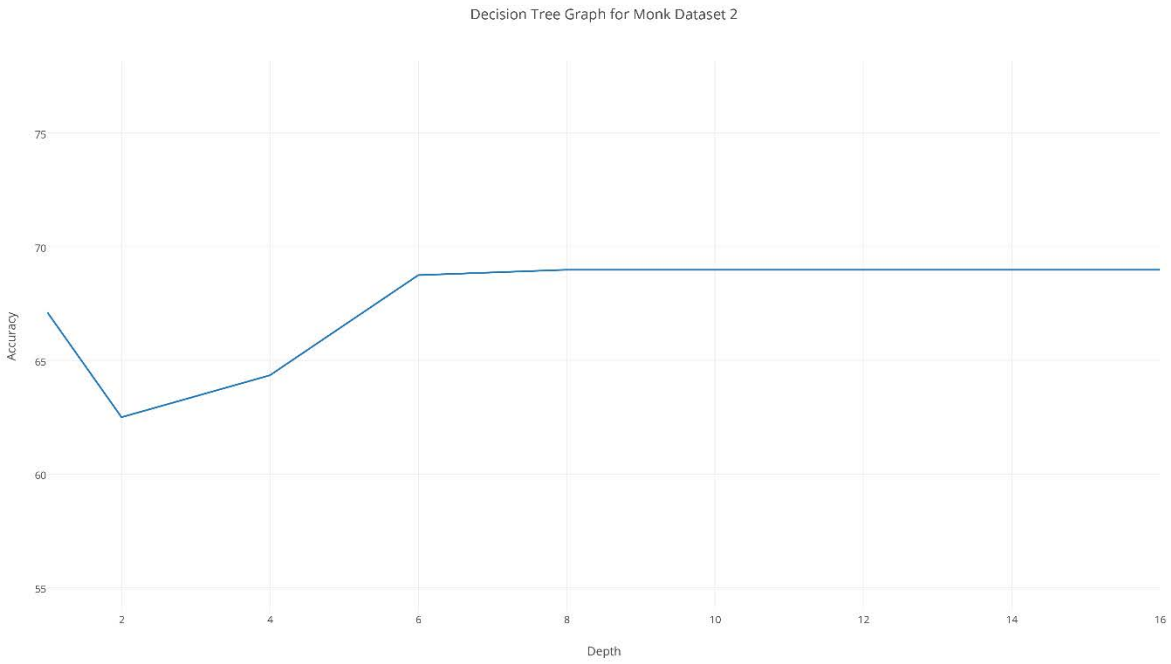
https://plot.ly/~mgalvank/3/



- Monk 2 dataset :

Depth	Error
1	0.32
2	0.37
4	0.35
6	0.31
8	0.31
10,12,14,16	0.31

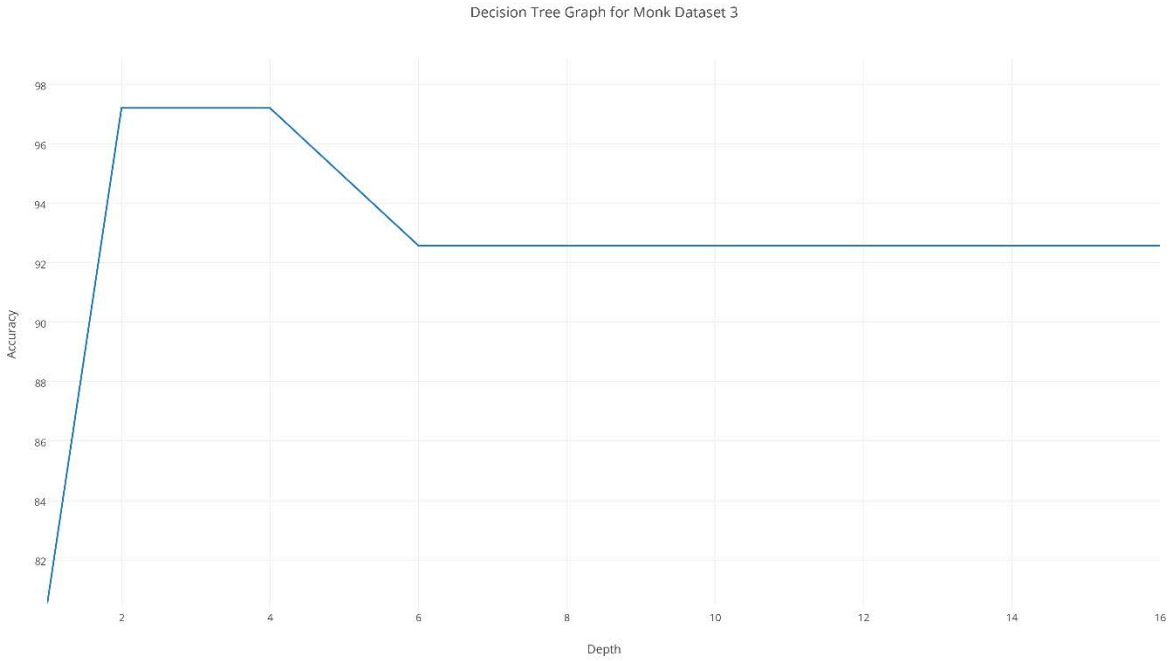
https://plot.ly/~mgalvank/0/



- Monk 3 dataset :

Depth	Error
1	0.19
2	0.027
4	0.074
6,8,…,16	0.074

https://plot.ly/~mgalvank/5/



2. Report the learned decision tree (depth 1 and depth 2) and report the confusion matrix for these two depths (A confusion matrix has the true label as rows and predicted labels in the columns. Each entry of the matrix is the number of examples. In a binary case, the top left corner is the number of negative examples correctly classified and the bottom right is the number of positives correctly classified).

- Monk 1 dataset
  a) Depth 1

  Tree : -

  If Feature a1 and Value 1 :

  Tree left->

  Result 0

  Tree right->

  Result 1

  Confusion matrix : -

```
Confusion Matrix :
|  N : 432     |   Predicted : No |   Predicted : Yes |
|--------------+------------------+-------------------|
| Actual : No  |               72 |               144 |
| Actual : Yes |               72 |               144 |
```

  b) Depth 2

  Tree : -

  If Feature a1 and Value 1.0 :

  Tree left->

  If Feature a2 and Value 1.0 :

  Tree left->

  Result 1.0

  Tree right->

  Result 0.0

  Tree right->

  If Feature a2 and Value 1.0 :

  Tree left->

  Result 0.0

  Tree right->

  Result 1.0

Confusion Matrix :-

N : 432	Predicted : No	Predicted : Yes
Actual : No	144	72
Actual : Yes	48	168

- Monk 2 dataset

a.) Depth 1

Tree :-

If Feature a4 and Value 1 :

Tree left->

Result 0

Tree right->

Result 0

Confusion Matrix :-

N : 432	Predicted : No	Predicted : Yes
Actual : No	290	0
Actual : Yes	142	0

b.) Depth 2

Tree :-

If Feature a4 and Value 1 :

Tree left->

If Feature a5 and Value 1 :

Tree left->

Result 0

Tree right->

Result 0

Tree right->

If Feature a5 and Value 2 :

Tree left->

Result 1

Tree right->

Result 0

Confusion Matrix :-

N : 432	Predicted : No	Predicted : Yes
Actual : No	244	46
Actual : Yes	116	26

- Monk 3 dataset

a.) Depth 1

Tree :-

If Feature a2 and Value 3 :

Tree left->

Result 0

Tree right->

Result 1

Confusion Matrix :-

N : 432	Predicted : No	Predicted : Yes
Actual : No	132	72
Actual : Yes	12	216

b.) Depth 2

Tree :-

If Feature a2 and Value 3 :

Tree left->

If Feature a5 and Value 3 :

Tree left->

Result 0

Tree right->

Result 0

Tree right->

If Feature a5 and Value 4 :

Tree left->

Result 0

Tree right->

Result 1

Confusion Matrix

N : 432	Predicted : No	Predicted : Yes
Actual : No	204	0
Actual : Yes	12	216

3. Now, use Weka's default decision tree (J48) algorithm on this training set to learn a decision tree. Report the tree and the confusion matrix on the test set. Do not change the default parameters of Weka.

a.) Monk 1 dataset :

```
J48 pruned tree
------------------

a5 = 1: 1 (29.0)
a5 = 2: 0 (31.0/11.0)
a5 = 3
|   a6 = 1: 0 (13.0/3.0)
|   a6 = 2
|   |   a3 = 1: 1 (7.0/2.0)
|   |   a3 = 2: 0 (10.0/3.0)
a5 = 4
|   a1 = 1: 0 (14.0/1.0)
|   a1 = 2
|   |   a2 = 1: 0 (6.0)
|   |   a2 = 2: 1 (4.0)
|   |   a2 = 3: 0 (1.0)
|   a1 = 3
|   |   a2 = 1: 1 (0.0)
|   |   a2 = 2: 0 (3.0)
|   |   a2 = 3: 1 (6.0)

Number of Leaves  :     12

Size of the tree :      18

=== Confusion Matrix ===

   a   b   <-- classified as
 186  30 |   a = 0
  75 141 |   b = 1
```

b.) Monk 2 dataset

```
J48 pruned tree
------------------

a4 = 1: 0 (54.0/15.0)
a4 = 2
|   a5 = 1
|   |   a3 = 1: 0 (7.0/1.0)
|   |   a3 = 2: 1 (5.0)
|   a5 = 2
|   |   a3 = 1
|   |   |   a6 = 1: 0 (3.0/1.0)
|   |   |   a6 = 2: 1 (4.0)
|   |   a3 = 2
|   |   |   a2 = 1: 1 (2.0)
|   |   |   a2 = 2: 0 (3.0)
|   |   |   a2 = 3: 0 (2.0)
|   a5 = 3: 0 (17.0/6.0)
|   a5 = 4: 0 (11.0/3.0)
a4 = 3
|   a3 = 1
|   |   a5 = 1: 0 (7.0/1.0)
|   |   a5 = 2: 1 (7.0/1.0)
|   |   a5 = 3: 1 (9.0/4.0)
|   |   a5 = 4
|   |   |   a2 = 1: 0 (2.0)
|   |   |   a2 = 2: 1 (3.0/1.0)
|   |   |   a2 = 3: 1 (2.0)
|   a3 = 2
|   |   a6 = 1
|   |   |   a1 = 1: 1 (4.0/1.0)
|   |   |   a1 = 2: 0 (4.0/1.0)
|   |   |   a1 = 3: 1 (4.0/1.0)
|   |   a6 = 2: 0 (19.0/4.0)

Number of Leaves  :     20

Size of the tree :      31

=== Confusion Matrix ===

   a   b   <-- classified as
 233  57 |   a = 0
  94  48 |   b = 1
```

c.) Monk 3 dataset

```
J48 pruned tree
------------------

a2 = 1
|   a5 = 1: 1 (12.0)
|   a5 = 2: 1 (9.0)
|   a5 = 3: 1 (6.0/1.0)
|   a5 = 4: 0 (12.0)
a2 = 2
|   a5 = 1: 1 (10.0)
|   a5 = 2: 1 (13.0/1.0)
|   a5 = 3: 1 (12.0/3.0)
|   a5 = 4: 0 (7.0)
a2 = 3: 0 (41.0/3.0)

Number of Leaves  :     9

Size of the tree :      12

=== Confusion Matrix ===

   a   b   <-- classified as
 204   0 |   a = 0
  12 216 |   b = 1
```

4. Repeat steps 2 and 3 with your “own” data set and report the confusion matrices. The total points on this homework is 20.

About the data set :-

Link : - https://archive.ics.uci.edu/ml/machine-learning-databases/spect/

SPECT heart data

-- Donors: Lukasz A.Kurgan, Krzysztof J. Cios

-- Date: 10/01/01

Description :-

The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature pattern was created for each patient. The pattern was further processed to obtain 22 binary feature patterns.

5. Number of Instances: 267

6. Number of Attributes: 23 (22 binary + 1 binary class)

7. Attribute Information:

1. OVERALL_DIAGNOSIS: 0,1 (class attribute, binary)
2. A1: 0,1 (the partial diagnosis 1, binary)
3. A2: 0,1 (the partial diagnosis 2, binary)
4. A3: 0,1 (the partial diagnosis 3, binary)
5. A4: 0,1 (the partial diagnosis 4, binary)
6. A5: 0,1 (the partial diagnosis 5, binary)
7. A6: 0,1 (the partial diagnosis 6, binary)
8. A7: 0,1 (the partial diagnosis 7, binary)
9. A8: 0,1 (the partial diagnosis 8, binary)
10. A9: 0,1 (the partial diagnosis 9, binary)
11. A10: 0,1 (the partial diagnosis 10, binary)
12. A11: 0,1 (the partial diagnosis 11, binary)
13. A12: 0,1 (the partial diagnosis 12, binary)
14. A13: 0,1 (the partial diagnosis 13, binary)
15. A14: 0,1 (the partial diagnosis 14, binary)
16. A15: 0,1 (the partial diagnosis 15, binary)
17. A16: 0,1 (the partial diagnosis 16, binary)
18. A17: 0,1 (the partial diagnosis 17, binary)
19. A18: 0,1 (the partial diagnosis 18, binary)
20. A19: 0,1 (the partial diagnosis 19, binary)
21. A20: 0,1 (the partial diagnosis 20, binary)
22. A21: 0,1 (the partial diagnosis 21, binary)
23. A22: 0,1 (the partial diagnosis 22, binary)

-- dataset is divided into:

    -- training data ("SPECT.train" 80 instances)

    -- testing data ("SPECT.test" 187 instances)

8. Missing Attribute Values: None

Depth 1

Tree :

If Feature 13 and Value 1.0 :

Tree left->

Result 1.0

Tree right->

Result 0.0

Confusion Matrix :

N : 187	Predicted : No	Predicted : Yes
Actual : No	13	2
Actual : Yes	70	102

Depth 2

Tree

If Feature 13 and Value 1.0 :

Tree left->

If Feature 8 and Value 0.0 :

Tree left->

Result 1.0

Tree right->

Result 1.0

Tree right->

If Feature 11 and Value 1.0 :

Tree left->

Result 1.0

Tree right->

Result 0.0

Confusion Matrix

N : 187	Predicted : No	Predicted : Yes
Actual : No	12	3
Actual : Yes	51	121

Weka

Tree : -

```
J48 pruned tree
------------------

a16 = 0
|   a8 = 0
|   |   a11 = 0: 0 (43.0/9.0)
|   |   a11 = 1: 1 (9.0/2.0)
|   a8 = 1
|   |   a3 = 0: 1 (4.0)
|   |   a3 = 1
|   |   |   a13 = 0: 0 (2.0)
|   |   |   a13 = 1: 1 (8.0/1.0)
a16 = 1: 1 (14.0/1.0)

Number of Leaves  :     6

Size of the tree :      11
```

Confusion Matrix

```
=== Confusion Matrix ===

   a   b   <-- classified as
  11   4 |   a = 0
  42 130 |   b = 1
```