Section B

Monday, January 23, 2023 8:04 PM

hw2_recit...

Recitation 2
Decision Trees

10-301/10-601: Introduction to Machine Learning
01/27/2023

1 Programming: Tree Structures and Algorithms

Topics Covered:

- Depth of nodes and trees
- Recursive traversal of trees
  - Depth First Search
    - Pre-order Traversal
    - In-order Traversal
    - Post-order Traversal
  - Breadth First Search (Self Study)
- Debugging in Python

Questions:

1. Depth of a tree definition

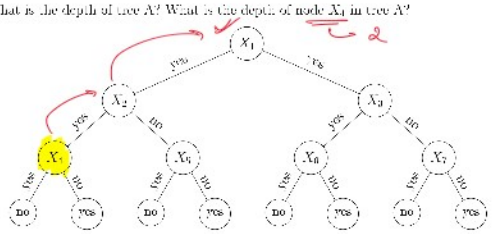
   # of layers till you reach the last leaf node.

   ⇒ Length of the longest path from the root node to a leaf node.

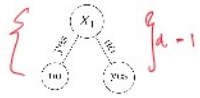
2. Depth of a node definition

   ⇒ distance from root : # of edges.

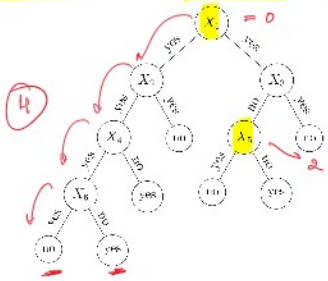
3. What is the depth of tree A? What is the depth of node X_4 in tree A?



4. What is the depth of tree B?



5. What is the depth of tree C? What are the depths of nodes X_1 and X_6 in tree C?



6. In-class coding and explanation of Depth First Traversal in Python.
Link to the code:
https://colab.research.google.com/drive/1KypCp2tPDad4goHjLIFH4DqbBnM5CfCr?usp=sharing

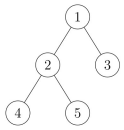
Pre-order, Inorder and Post-order Tree Traversal

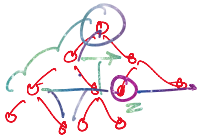
```
# This class represents an individual node
class Node:
    def __init__(self, key):
        self.left = None
        self.right = None
        self.val = key

def traversal1(root):
    if root is not None:
        # First recurse on left child
        traversal1(root.left)
        # then recurse on right child
        traversal1(root.right)
        # now print the data of node
        print(root.val, end = "\t")

def traversal2(root):
    if root is not None:
        # First print the data of node
        print(root.val, end = "\t")
        # Then recurse on left child
        traversal2(root.left)
        # Finally recurse on right child
        traversal2(root.right)

def traversal3(root):
    if root is not None:
        # First recurse on left child
        traversal3(root.left)
        # then print the data of node
        print(root.val, end="\t")
        # now recurse on right child
        traversal3(root.right)

def build_a_tree():
    root = Node(1)
    root.left = Node(2)
    root.right = Node(3)
    root.left.left = Node(4)
    root.left.right = Node(5)
    return root
```

```
if __name__ == '__main__':
    root = build_a_tree()
    print('traversal1 of the binary tree is: ')
    traversal1(root)
    print()
    print('traversal2 of the binary tree is: ')
    traversal2(root)
    print()
    print('traversal3 of the binary tree is: ')
    traversal3(root)
```

Now, identify which traversal function is pre-order, in-order, post-order DFS:

- traversal1() is
- traversal2() is
- traversal3() is





Code Output

```
traversal1 of the binary tree is:


traversal2 of the binary tree is


traversal3 of the binary tree is
```


2 ML Concepts: Mutual Information

Information Theory Definitions:

- $H(Y) = -\sum_{y \in values(Y)} P(Y = y) \log_2 P(Y = y)$
- $H(Y \mid X = x) = -\sum_{y \in values(Y)} P(Y = y | X = x) \log_2 P(Y = y | X = x)$
- $H(Y \mid X) = \sum_{x \in values(X)} P(X = x) H(Y \mid X = x)$
- $I(X; Y) = H(Y) - H(Y \mid X)$

Exercises

1. Calculate the entropy of tossing a fair coin.

$P(H) = P(T) = 0.5$

$-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

2. Calculate the entropy of tossing a coin that lands only on tails. *Note:* $0 \cdot \log_2(0) = 0$.

$P(H) = 0 \quad P(T) = 1$

$-0 \log 0 - 1 \log_2 1 = 0$

3. Calculate the entropy of a fair dice roll.

$-\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6}$

4. When is the mutual information $I(X; Y) = 0$?

$H(Y) - H(Y|X) = 0$

$H(Y) = H(Y|X)$

iff $X \perp\!\!\!\perp Y$

Used in Decision Trees: $I(Y; X) = H(Y) - H(Y|X)$

Outlook (X_1)	Temperature (X_2)	Humidity (X_3)	Play Tennis? (Y)
sunny	hot	high	no
overcast	hot	high	yes
rain	mild	high	yes
rain	cool	normal	yes
sunny	mild	high	no
sunny	mild	normal	yes
rain	mild	normal	yes
overcast	hot	normal	yes

1. Using the dataset above, calculate the mutual information for each feature (X_1, X_2, X_3) to determine the root node for a Decision Tree trained on the above data.
   - What is $I(Y; X_1)$?
   - What is $I(Y; X_2)$?
   - What is $I(Y; X_3)$?
   - Which feature should be split on at the root node?

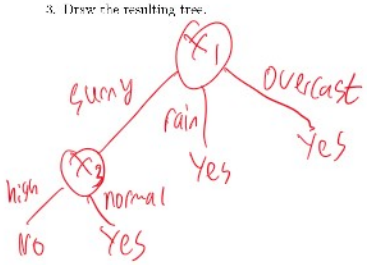
   X_1

2. Calculate what the next split should be.

   $I(Y; X_2) \approx 0.25$

   $I(Y; X_3) \approx 0.918$

3. Draw the resulting tree.



$H(Y) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} \approx 0.811$

$H(Y|X_1) = P(X_1=\text{sunny})\,H(Y|X_1=\text{sun}) = \frac{3}{8} \cdot 0.918 \approx 0.344$

$+ P(X_1=\text{rain})\,H(Y|X_1=\text{rain})$ 0

$+ P(X_1=\text{overcast})\,H(Y|X_1=\text{overcast})$ 0

$I(Y; X_1) = 0.811 - 0.344 = 0.467$

$I(Y; X_2) = 0.061$

$I(Y; X_3) = 0.311$

3 ML Concepts: Construction of Decision Trees

In this section, we will go over how to construct our decision tree learner on a high level. The following questions will help guide the discussion.

1. What exactly are the tasks we are tackling?
2. What are the inputs and outputs at training time? At testing time?
3. At each node of the tree, what do we need to store?
4. What do we need to do at training time?
5. What happens if max depth is 0?
6. What happens if max depth is greater than the number of attributes?

1) The task: Given a set of train data, test data, max-depth, we want:
   a) use train data to learn decision tree classifier
   b) use the trained classifier to predict labels for both train and test dataset
   c) calculate error for both train and test data

2)

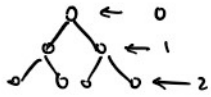
	Train	Test
inputs	• train data • max-depth of tree	• test data (must be provided in the same format as train data)
Outputs	• fully trained decision tree classifier	• predictions for every entry/row in test dataset

3) class Node:

```
def __init__(self, m):
    self.m = m
    self.n = n
    self.a = a
    self.b = b
    ⋮
```

- attribute that we split on
- left and right child nodes
- depth of node
- subset of data used at that given node

4) • consider "stopping criterion"
   → max-depth reached
   → node is pure (data is all one label, entropy = 0)
   ↳ simple majority vote
   • Calculate entropy and mutual information for only non-used attributes, and select best attribute to split on.
   • Split the data according to the best attribute

5) 

0

majority vote

6) depth of tree $\leq \min(\#\text{ attributes}, \text{max-depth})$

stop growing tree after all attributes have been used

4 Programming: Debugging with Trees

pdb and common commands

- `import pdb; pdb.set_trace()` (`breakpoint()` also allowed as per PEP 553)
- `p variable` (print value of variable)
- `n` (next)
- `s` (step into subroutine)
- `ENTER` (repeat previous command)
- `q` (quit)
- `l` (list where you are)
- `b` (breakpoint)
- `c` (continue)
- `r` (continue until the end of the subroutine)
- `!code` (run Python code)

Real Practice

These are some (contrived) examples based on actual bugs previous students had. Link to the code: https://colab.research.google.com/drive/1KypCp2tPDad4gdHjL1FH4DqbBnM5CfCr?usp=sharing

Buggy Code

```
# Reverse the rows of a 2D array
def reverse_rows(original):
    rows = len(original)
    cols = len(original[0])

    new = [[0] * cols] * rows

    for i in range(rows):
        for j in range(cols):
            new_index = rows - i
            new[new_index][j] = original[i][j]

    return new


if __name__ == '__main__':
    a = [[1, 2],
         [3, 4],
         [5, 6]]
    print(reverse_rows(a))
```


Buggy Code

```
import numpy as np

# biggest_col takes a binary 2D array and returns the index of the
# column with the most non-zero values. In case of a tie, return
# the smallest index.
def biggest_col(mat):
    num_col = len(mat[0])
    max_count = -1
    max_index = -1

    # iterate over the columns of the matrix
    for col in range(num_col):
        # counts the number of nonzero values
        count = np.count_nonzero(mat[:, col])
        # change max if needed
        if count >= max_count:
            max_count = count
            max_index = col

    return max_index


# Helper function that returns the number of nonzero elements in
# mat in column col.
def get_count(mat, col):
    num_row = len(mat)
    count = 0
    for row in range(num_row):
        count += (mat[row][col] == 0)
    return count


if __name__ == '__main__':
    # Expected answer: column index 2
    mat = [[1, 0, 0, 1],
           [0, 1, 1, 1],
           [1, 0, 0, 0],
           [0, 1, 1, 1],
           [0, 0, 1, 0]]
    assert biggest_col(mat) == 2
```