CSE151A_PA2

March 18, 2021

[1]:

```python
import numpy as np
import pandas as pd
import math
from scipy.stats import entropy
import scipy.stats
```

[2]:

```python
train = np.loadtxt('data/pa2train.txt')
test = np.loadtxt('data/pa2test.txt')
valid = np.loadtxt('data/pa2validation.txt')
features = np.loadtxt('data/pa2features.txt',dtype = 'object')
```

[3]:

```python
Xtrain = train
ytrain = train[:,-1]

Xtest = test
ytest = test[:,-1]

Xvalid = valid
yvalid = valid[:,-1]
```

1. First, build an ID3 Decision Tree classifier based on the data in pa2train.txt. Do not use pruning. Draw the first three levels decision tree that you obtain. For each node that you draw, if it is a leaf node, write down the label that will be predicted for this node, as well as how many of the training data points lie in this node. If it is an internal node, write down the splitting rule for the node, as well as how many of the training data points lie in this node. (Hint: If your code is correct, the root node will involve the rule Feature 5 < 0.5.)

[4]:

```python
class Node:
    def __init__(self):
        self.data = Node
        self.yesBranch = None
        self.noBranch = None
        self.label = None
        self.decision = None
        self.indices = None
        self.isLeaf = None
        self.isPure = None
        self.numPoints = None
        self.feature = None
        self.thresh = None
```

[5]:

```
#entropy calculation helper function
def calcEntropy(data):
    N = len(data)
    #get num of yes/no labels
    yes_cnt = np.sum(data[:,-1]==1)
    no_cnt = np.sum(data[:,-1] ==0)
    if yes_cnt ==0 or no_cnt ==0:
        return 0
    return ((yes_cnt/N)*np.log(yes_cnt/N)+ (no_cnt/N)*np.log(no_cnt/N)) *-1
```

[6]:

```
#helper function
#check is a node is pure
def isPure(node):
    if len(np.unique(node.data[:,-1])) !=1:
        return False
    else:
        return True
```

[7]:

```
#builds tree given training data
def ID3DecisionTree(data):
    queue = []
    rootNode = Node()
    rootNode.data = data
    queue.append(rootNode)
    #while queue is not empty
    while len(queue) >0:
        node = queue.pop(0)
        #only add node to queue if impure
        if isPure(node):
            # if pure set label
            node.label= node.data[0][-1]
        else:
            node1,node2 = splitRule(node)
            queue.append(node1)
            queue.append(node2)

    return rootNode
```

[8]:

```
# determines best feature/threshold to split node at
def splitRule(node):
    data = node.data
    N = len(data)
    minEntropy = np.inf
    #for each feature
    for i in range(len(features)):
        #sort data by i'th column/feature
        sorted_data = data[data[:,i].argsort()]
        #for each feature vector
        for x in range(len(sorted_data)-1):
            #skip equivalent values
            if sorted_data[x][i] == sorted_data[x+1][i]:
                continue
            #compute midpoint between adjacent sorted values within ith feature
            threshold = (sorted_data[x][i] + sorted_data[x+1][i])/2
            ent = (((x+1)/N))*calcEntropy(sorted_data[:x+1]) + (1-((x+1)/N))*calcEntropy(sorted_data[x+1:])
            if ent < minEntropy:
                minEntropy = ent
                bestFeature = i
                bestThresh = threshold
                best_x_i = x

    node.feature = bestFeature
    node.thresh = bestThresh
    print("split at x[" +str(bestFeature)+ "] <= " +str(bestThresh))
    node.yesBranch = Node()
    node.noBranch = Node()
    sorted_d = data[data[:,bestFeature].argsort()]
    node.yesBranch.data = sorted_d[:best_x_i+1]
    print("num points in yes branch ",len(node.yesBranch.data))
    node.noBranch.data = sorted_d[best_x_i+1:]
    print("num points in no branch ",len(node.noBranch.data))
    print("next node \u2193 \n")
    return node.yesBranch, node.noBranch
```

[9]:

```
#traverse tree at a node to get label of a feature vector
def getLabel(node, datapoint):
    #keep traversing until node has label
    while node.label is None:
        #feature vector at i'th feature is less than threshold
        x_f_i = datapoint[node.feature]
        #visit left yes branch if datapoint at feature is less than threshold
        #else visit no branch
        if x_f_i < node.thresh:
            node = node.yesBranch
        else:
            node = node.noBranch

    return node.label
```

[10]:

```
# while impure:
# if np.sum(ytrain[node1.indices] == 0) == 0:
#         node1.label = 1
#         node1.isPure = True
#     elif np.sum(ytrain[node1.indices] == 1) == 0:
#         node1.label = 0
#         node1.isPure = True
```

[ ]:

```

```

[11]:

```
#helper function for pruning, computes mode
def mode(data):
    N = len(data)
    data = np.array(data)
    #compute mode of labels
    label_mode = scipy.stats.mode(data[:,-1])
    label = label_mode[0][0]
    #computes error of most common label
    err = 1 - (scipy.stats.mode(data[:,-1])[1][0] / N)
    return int(label), err
```

[12]:

```
#a bit confusing to interpret based on these print statements
tree = ID3DecisionTree(train)
```

```
split at x[4] <= 0.5
num points in yes branch  1319
num points in no branch  681
next node ↓

split at x[0] <= 415000.0
num points in yes branch  1284
num points in no branch  35
next node ↓

split at x[4] <= 1.5
num points in yes branch  292
num points in no branch  389
next node ↓

split at x[16] <= 2506.5
num points in yes branch  704
num points in no branch  580
next node ↓

split at x[20] <= 208.0
num points in yes branch  4
num points in no branch  31
```

```
next node ↓

split at x[19] <= 584.5
num points in yes branch  134
num points in no branch  158
next node ↓

split at x[20] <= 2006.0
num points in yes branch  232
num points in no branch  157
next node ↓

split at x[0] <= 75000.0
num points in yes branch  393
num points in no branch  311
next node ↓

split at x[0] <= 25000.0
num points in yes branch  9
num points in no branch  571
next node ↓

split at x[16] <= 2174.0
num points in yes branch  9
num points in no branch  22
next node ↓

split at x[11] <= 231.5
num points in yes branch  54
num points in no branch  80
next node ↓

split at x[11] <= 1461.0
num points in yes branch  22
num points in no branch  136
next node ↓

split at x[18] <= 2476.0
num points in yes branch  182
num points in no branch  50
next node ↓

split at x[18] <= 13075.0
num points in yes branch  147
num points in no branch  10
next node ↓

split at x[12] <= 46620.5
```

```
num points in yes branch  349
num points in no branch  44
next node ↓

split at x[0] <= 115000.0
num points in yes branch  35
num points in no branch  276
next node ↓

split at x[18] <= 750.0
num points in yes branch  4
num points in no branch  5
next node ↓

split at x[17] <= 14935.5
num points in yes branch  532
num points in no branch  39
next node ↓

split at x[21] <= 2121.5
num points in yes branch  3
num points in no branch  19
next node ↓

split at x[3] <= 2.5
num points in yes branch  53
num points in no branch  1
next node ↓

split at x[11] <= 316.0
num points in yes branch  3
num points in no branch  77
next node ↓

split at x[1] <= 1.5
num points in yes branch  10
num points in no branch  12
next node ↓

split at x[20] <= 1911.5
num points in yes branch  84
num points in no branch  52
next node ↓

split at x[18] <= 2426.0
num points in yes branch  181
num points in no branch  1
next node ↓
```

```
split at x[18] <= 13894.0
num points in yes branch  1
num points in no branch  9
next node ↓

split at x[0] <= 25000.0
num points in yes branch  45
num points in no branch  304
next node ↓

split at x[1] <= 1.5
num points in yes branch  14
num points in no branch  30
next node ↓

split at x[10] <= 668.5
num points in yes branch  11
num points in no branch  24
next node ↓

split at x[0] <= 125000.0
num points in yes branch  30
num points in no branch  246
next node ↓

split at x[11] <= 204348.0
num points in yes branch  499
num points in no branch  33
next node ↓

split at x[0] <= 475000.0
num points in yes branch  5
num points in no branch  14
next node ↓

split at x[0] <= 190000.0
num points in yes branch  25
num points in no branch  28
next node ↓

split at x[17] <= 6372.5
num points in yes branch  72
num points in no branch  5
next node ↓

split at x[12] <= 701.5
num points in yes branch  9
num points in no branch  3
next node ↓

split at x[18] <= 11929.0
num points in yes branch  51
num points in no branch  1
next node ↓

split at x[6] <= 1.0
num points in yes branch  66
num points in no branch  115
next node ↓

split at x[21] <= 646.0
num points in yes branch  33
num points in no branch  12
next node ↓

split at x[11] <= 16815.5
num points in yes branch  137
num points in no branch  167
next node ↓

split at x[10] <= 58843.5
num points in yes branch  10
num points in no branch  4
next node ↓

split at x[10] <= 50823.5
num points in yes branch  12
num points in no branch  18
next node ↓

split at x[5] <= 1.0
num points in yes branch  7
num points in no branch  4
next node ↓

split at x[16] <= 1340.0
num points in yes branch  10
num points in no branch  14
next node ↓

split at x[21] <= 2888.5
num points in yes branch  195
num points in no branch  51
next node ↓
```

```
split at x[10] <= 146305.0
num points in yes branch  454
num points in no branch  45
next node ↓

split at x[18] <= 14104.5
num points in yes branch  4
num points in no branch  1
next node ↓

split at x[18] <= 63135.0
num points in yes branch  12
num points in no branch  2
next node ↓

split at x[2] <= 0.5
num points in yes branch  3
num points in no branch  25
next node ↓

split at x[2] <= 0.5
num points in yes branch  19
num points in no branch  53
next node ↓

split at x[0] <= 225000.0
num points in yes branch  4
num points in no branch  1
next node ↓

split at x[14] <= 903.5
num points in yes branch  4
num points in no branch  5
next node ↓

split at x[3] <= 1.5
num points in yes branch  24
num points in no branch  27
next node ↓

split at x[0] <= 65000.0
num points in yes branch  74
num points in no branch  41
next node ↓

split at x[17] <= 1348.5
num points in yes branch  19
num points in no branch  14
```

```
next node ↓

split at x[16] <= 1583.5
num points in yes branch  9
num points in no branch  3
next node ↓

split at x[21] <= 1514.0
num points in yes branch  107
num points in no branch  30
next node ↓

split at x[3] <= 2.5
num points in yes branch  166
num points in no branch  1
next node ↓

split at x[12] <= 49649.5
num points in yes branch  6
num points in no branch  4
next node ↓

split at x[3] <= 1.5
num points in yes branch  6
num points in no branch  12
next node ↓

split at x[0] <= 85000.0
num points in yes branch  1
num points in no branch  3
next node ↓

split at x[16] <= 1919.0
num points in yes branch  10
num points in no branch  4
next node ↓

split at x[10] <= 19215.0
num points in yes branch  180
num points in no branch  15
next node ↓

split at x[0] <= 190000.0
num points in yes branch  9
num points in no branch  42
next node ↓

split at x[17] <= 10835.5
```

```
num points in yes branch  443
num points in no branch  11
next node ↓

split at x[18] <= 4502.5
num points in yes branch  10
num points in no branch  35
next node ↓

split at x[0] <= 290000.0
num points in yes branch  2
num points in no branch  1
next node ↓

split at x[21] <= 839.5
num points in yes branch  22
num points in no branch  3
next node ↓

split at x[0] <= 25000.0
num points in yes branch  6
num points in no branch  13
next node ↓

split at x[14] <= 15900.0
num points in yes branch  32
num points in no branch  21
next node ↓

split at x[20] <= 967.5
num points in yes branch  2
num points in no branch  2
next node ↓

split at x[16] <= 8400.0
num points in yes branch  26
num points in no branch  1
next node ↓

split at x[18] <= 1867.0
num points in yes branch  64
num points in no branch  10
next node ↓

split at x[3] <= 1.5
num points in yes branch  17
num points in no branch  24
next node ↓
```

```
split at x[11] <= 7776.0
num points in yes branch  7
num points in no branch  12
next node ↓

split at x[10] <= 585.0
num points in yes branch  2
num points in no branch  7
next node ↓

split at x[16] <= 1641.0
num points in yes branch  95
num points in no branch  12
next node ↓

split at x[16] <= 2503.0
num points in yes branch  165
num points in no branch  1
next node ↓

split at x[15] <= 28057.0
num points in yes branch  2
num points in no branch  2
next node ↓

split at x[10] <= 59051.0
num points in yes branch  3
num points in no branch  3
next node ↓

split at x[0] <= 55000.0
num points in yes branch  1
num points in no branch  11
next node ↓

split at x[12] <= 23252.0
num points in yes branch  6
num points in no branch  4
next node ↓

split at x[0] <= 185000.0
num points in yes branch  52
num points in no branch  128
next node ↓

split at x[19] <= 7026.0
num points in yes branch  7
num points in no branch  2
next node ↓

split at x[10] <= 49829.0
num points in yes branch  40
num points in no branch  2
next node ↓

split at x[20] <= 809.5
num points in yes branch  95
num points in no branch  348
next node ↓

split at x[21] <= 3544.5
num points in yes branch  5
num points in no branch  6
next node ↓

split at x[14] <= 95643.0
num points in yes branch  4
num points in no branch  6
next node ↓

split at x[18] <= 9000.0
num points in yes branch  29
num points in no branch  6
next node ↓

split at x[0] <= 255000.0
num points in yes branch  1
num points in no branch  1
next node ↓

split at x[0] <= 220000.0
num points in yes branch  1
num points in no branch  2
next node ↓

split at x[7] <= 1.0
num points in yes branch  9
num points in no branch  4
next node ↓

split at x[18] <= 2555.5
num points in yes branch  12
num points in no branch  9
next node ↓
```

```
split at x[14] <= 29380.5
num points in yes branch  10
num points in no branch  16
next node ↓

split at x[1] <= 1.5
num points in yes branch  6
num points in no branch  4
next node ↓

split at x[13] <= 455.5
num points in yes branch  2
num points in no branch  22
next node ↓

split at x[21] <= 97.5
num points in yes branch  7
num points in no branch  5
next node ↓

split at x[11] <= 3124.0
num points in yes branch  4
num points in no branch  3
next node ↓

split at x[10] <= 46792.5
num points in yes branch  94
num points in no branch  1
next node ↓

split at x[13] <= 5583.0
num points in yes branch  8
num points in no branch  4
next node ↓

split at x[13] <= 25530.0
num points in yes branch  74
num points in no branch  91
next node ↓

split at x[10] <= 59905.5
num points in yes branch  2
num points in no branch  1
next node ↓

split at x[11] <= 12973.0
num points in yes branch  3
num points in no branch  3
```

```
next node ↓

split at x[21] <= 1189.0
num points in yes branch  41
num points in no branch  11
next node ↓

split at x[19] <= 1728.0
num points in yes branch  106
num points in no branch  22
next node ↓

split at x[2] <= 0.5
num points in yes branch  1
num points in no branch  6
next node ↓

split at x[21] <= 10234.5
num points in yes branch  24
num points in no branch  16
next node ↓

split at x[16] <= 3153.5
num points in yes branch  18
num points in no branch  77
next node ↓

split at x[10] <= 1809.5
num points in yes branch  22
num points in no branch  326
next node ↓

split at x[0] <= 165000.0
num points in yes branch  1
num points in no branch  4
next node ↓

split at x[3] <= 1.5
num points in yes branch  3
num points in no branch  1
next node ↓

split at x[15] <= 194572.5
num points in yes branch  28
num points in no branch  1
next node ↓

split at x[8] <= 1.0
```

```
num points in yes branch  4
num points in no branch  2
next node ↓

split at x[0] <= 250000.0
num points in yes branch  8
num points in no branch  1
next node ↓

split at x[20] <= 890.5
num points in yes branch  4
num points in no branch  8
next node ↓

split at x[0] <= 45000.0
num points in yes branch  8
num points in no branch  2
next node ↓

split at x[0] <= 40000.0
num points in yes branch  3
num points in no branch  1
next node ↓

split at x[0] <= 335000.0
num points in yes branch  21
num points in no branch  1
next node ↓

split at x[12] <= 4860.0
num points in yes branch  3
num points in no branch  4
next node ↓

split at x[10] <= 7900.5
num points in yes branch  1
num points in no branch  4
next node ↓

split at x[10] <= 13850.5
num points in yes branch  2
num points in no branch  1
next node ↓

split at x[20] <= 198.0
num points in yes branch  46
num points in no branch  48
next node ↓
```

```
split at x[14] <= 1027.0
num points in yes branch  4
num points in no branch  4
next node ↓

split at x[15] <= 21960.5
num points in yes branch  51
num points in no branch  23
next node ↓

split at x[21] <= 656.5
num points in yes branch  33
num points in no branch  8
next node ↓

split at x[15] <= 2474.5
num points in yes branch  95
num points in no branch  11
next node ↓

split at x[11] <= 4137.5
num points in yes branch  18
num points in no branch  6
next node ↓

split at x[14] <= 10768.0
num points in yes branch  9
num points in no branch  7
next node ↓

split at x[9] <= 1.0
num points in yes branch  10
num points in no branch  8
next node ↓

split at x[18] <= 35025.5
num points in yes branch  74
num points in no branch  3
next node ↓

split at x[15] <= 4690.5
num points in yes branch  12
num points in no branch  10
next node ↓

split at x[20] <= 4002.5
num points in yes branch  214
num points in no branch  112
next node ↓

split at x[12] <= 36038.5
num points in yes branch  7
num points in no branch  1
next node ↓

split at x[20] <= 2500.0
num points in yes branch  5
num points in no branch  3
next node ↓

split at x[10] <= 19641.0
num points in yes branch  1
num points in no branch  2
next node ↓

split at x[7] <= 4.5
num points in yes branch  20
num points in no branch  1
next node ↓

split at x[2] <= 0.5
num points in yes branch  1
num points in no branch  2
next node ↓

split at x[5] <= 1.0
num points in yes branch  41
num points in no branch  5
next node ↓

split at x[21] <= 1206.5
num points in yes branch  45
num points in no branch  3
next node ↓

split at x[12] <= 825.5
num points in yes branch  2
num points in no branch  2
next node ↓

split at x[14] <= 21774.5
num points in yes branch  2
num points in no branch  21
next node ↓
```

```
split at x[14] <= 16077.5
num points in yes branch  31
num points in no branch  2
next node ↓

split at x[0] <= 155000.0
num points in yes branch  4
num points in no branch  4
next node ↓

split at x[0] <= 205000.0
num points in yes branch  41
num points in no branch  54
next node ↓

split at x[10] <= 10893.0
num points in yes branch  8
num points in no branch  3
next node ↓

split at x[3] <= 1.5
num points in yes branch  4
num points in no branch  2
next node ↓

split at x[0] <= 330000.0
num points in yes branch  7
num points in no branch  2
next node ↓

split at x[13] <= 98920.5
num points in yes branch  6
num points in no branch  1
next node ↓

split at x[12] <= 373.5
num points in yes branch  1
num points in no branch  9
next node ↓

split at x[13] <= 8297.5
num points in yes branch  2
num points in no branch  6
next node ↓

split at x[5] <= 1.0
num points in yes branch  73
num points in no branch  1
```

```
next node ↓

split at x[3] <= 1.5
num points in yes branch  2
num points in no branch  1
next node ↓

split at x[11] <= 6472.0
num points in yes branch  4
num points in no branch  6
next node ↓

split at x[20] <= 3324.5
num points in yes branch  181
num points in no branch  33
next node ↓

split at x[5] <= 2.5
num points in yes branch  4
num points in no branch  1
next node ↓

split at x[8] <= 1.0
num points in yes branch  11
num points in no branch  9
next node ↓

split at x[14] <= -490.0
num points in yes branch  1
num points in no branch  40
next node ↓

split at x[6] <= 1.0
num points in yes branch  2
num points in no branch  3
next node ↓

split at x[10] <= 9244.5
num points in yes branch  1
num points in no branch  2
next node ↓

split at x[13] <= 25477.5
num points in yes branch  20
num points in no branch  1
next node ↓

split at x[18] <= 343.0
```

```
num points in yes branch  22
num points in no branch  9
next node ↓

split at x[1] <= 1.5
num points in yes branch  2
num points in no branch  2
next node ↓

split at x[14] <= -354.0
num points in yes branch  1
num points in no branch  40
next node ↓

split at x[10] <= 7715.5
num points in yes branch  52
num points in no branch  2
next node ↓

split at x[10] <= 640.0
num points in yes branch  2
num points in no branch  6
next node ↓

split at x[2] <= 0.5
num points in yes branch  1
num points in no branch  6
next node ↓

split at x[17] <= 3835.5
num points in yes branch  5
num points in no branch  1
next node ↓

split at x[17] <= 2084.5
num points in yes branch  45
num points in no branch  28
next node ↓

split at x[11] <= 48957.0
num points in yes branch  5
num points in no branch  1
next node ↓

split at x[15] <= 38724.0
num points in yes branch  111
num points in no branch  70
next node ↓
```

```
split at x[15] <= 99997.0
num points in yes branch  23
num points in no branch  10
next node ↓

split at x[14] <= 29979.5
num points in yes branch  8
num points in no branch  3
next node ↓

split at x[21] <= 1359.5
num points in yes branch  39
num points in no branch  1
next node ↓

split at x[10] <= 1328.0
num points in yes branch  1
num points in no branch  2
next node ↓

split at x[0] <= 150000.0
num points in yes branch  11
num points in no branch  11
next node ↓

split at x[3] <= 1.5
num points in yes branch  26
num points in no branch  26
next node ↓

split at x[17] <= 3745.0
num points in yes branch  8
num points in no branch  20
next node ↓

split at x[0] <= 85000.0
num points in yes branch  26
num points in no branch  44
next node ↓

split at x[12] <= 93450.5
num points in yes branch  17
num points in no branch  6
next node ↓

split at x[2] <= 0.5
num points in yes branch  1
num points in no branch  7
next node ↓

split at x[14] <= 52494.0
num points in yes branch  2
num points in no branch  1
next node ↓

split at x[10] <= 2973.0
num points in yes branch  19
num points in no branch  20
next node ↓

split at x[0] <= 135000.0
num points in yes branch  3
num points in no branch  8
next node ↓

split at x[10] <= 1240.0
num points in yes branch  9
num points in no branch  2
next node ↓

split at x[10] <= 237.5
num points in yes branch  10
num points in no branch  16
next node ↓

split at x[0] <= 330000.0
num points in yes branch  24
num points in no branch  2
next node ↓

split at x[2] <= 0.5
num points in yes branch  2
num points in no branch  6
next node ↓

split at x[17] <= 500.0
num points in yes branch  2
num points in no branch  24
next node ↓

split at x[10] <= 4412.0
num points in yes branch  1
num points in no branch  16
next node ↓
```

```
split at x[13] <= 104732.5
num points in yes branch  5
num points in no branch  1
next node ↓

split at x[10] <= 785.5
num points in yes branch  13
num points in no branch  6
next node ↓

split at x[10] <= 2790.0
num points in yes branch  2
num points in no branch  1
next node ↓

split at x[0] <= 175000.0
num points in yes branch  5
num points in no branch  4
next node ↓

split at x[10] <= 50.5
num points in yes branch  9
num points in no branch  1
next node ↓

split at x[1] <= 1.5
num points in yes branch  6
num points in no branch  18
next node ↓

split at x[10] <= 94094.5
num points in yes branch  5
num points in no branch  1
next node ↓

split at x[0] <= 75000.0
num points in yes branch  21
num points in no branch  3
next node ↓

split at x[11] <= 767.5
num points in yes branch  2
num points in no branch  4
next node ↓

split at x[10] <= 512.5
num points in yes branch  3
num points in no branch  2
```

```
next node ↓

split at x[10] <= 195.0
num points in yes branch  3
num points in no branch  3
next node ↓

split at x[11] <= 1246.5
num points in yes branch  14
num points in no branch  4
next node ↓

split at x[15] <= 38969.0
num points in yes branch  1
num points in no branch  20
next node ↓

split at x[1] <= 1.5
num points in yes branch  1
num points in no branch  2
next node ↓

split at x[2] <= 0.5
num points in yes branch  2
num points in no branch  2
next node ↓
```

[13]:

```
from graphviz import Digraph
```

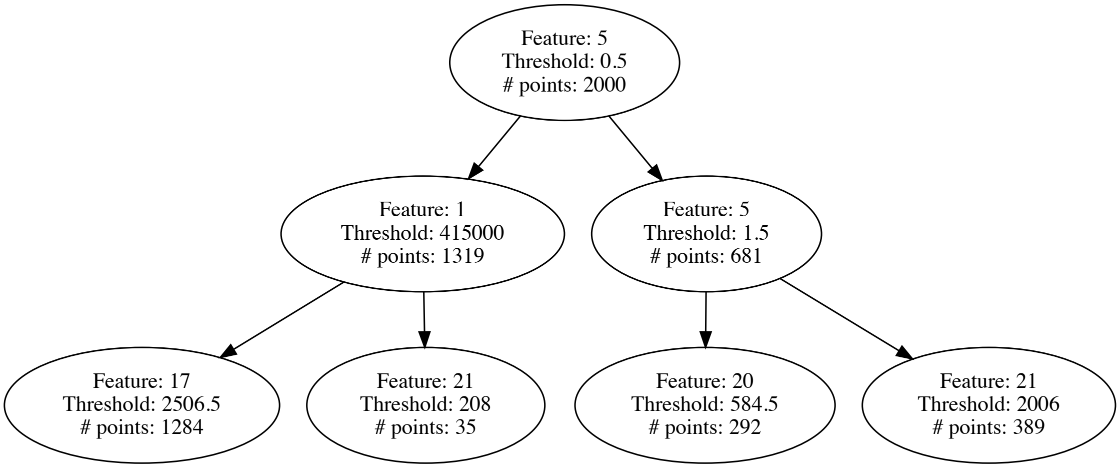
[14]:

```
dot = Digraph()
dot.attr(size='6,6')
dot.node('1','Feature: 5 \n Threshold: 0.5 \n # points: 2000', **{'width':'0.
 ↪5', 'height':'0.5'})
dot.node('2','Feature: 1 \n Threshold: 415000 \n # points: 1319', **{'width':'0.
 ↪5', 'height':'0.5'})
dot.node('3','Feature: 5 \n Threshold: 1.5 \n # points: 681', **{'width':'0.5',␣
 ↪'height':'0.5'})
dot.node('4','Feature: 17 \n Threshold: 2506.5 \n # points: 1284', **{'width':
 ↪'0.5', 'height':'0.5'})
dot.node('5','Feature: 21 \n Threshold: 208 \n # points: 35', **{'width':'0.5',␣
 ↪'height':'0.5'})
dot.node('6','Feature: 20 \n Threshold: 584.5 \n # points: 292',**{'width':'0.
 ↪5', 'height':'0.5'})
dot.node('7','Feature: 21 \n Threshold: 2006 \n # points: 389', **{'width':'0.
 ↪5', 'height':'0.5'})
```

```
dot.edges(['12'])
dot.edges(['13'])
dot.edges(['24'])
dot.edges(['25'])
dot.edges(['36'])
dot.edges(['37'])

dot
```

[14]:



2. What is the training and test error of your classifier in part (1), where test error is measured on the data in pa2test.txt?

[15]:

```
#get number of incorrectly predicted labels
def calcError(tree, data):
    correct = 0
    N = len(data)
    #loop through node data, each feature vector
    for xi in data:
        #gets predicted label
        label_xi = getLabel(tree, xi)
        #gets true label
        true_label = xi[-1]
        if true_label == label_xi:
            correct+=1
    #proportion of correct labels
    p = correct/N
    return 1 - p
```

[16]:

```
train_error = calcError(tree, train)
print("training error of id3 decision tree classifier: ", train_error)
```

```
training error of id3 decision tree classifier:  0.0
```

[17]:
```
test_error = calcError(tree, test)
print("test error of id3 decision tree classifier: ", test_error)
```

```
test error of id3 decision tree classifier:  0.17300000000000004
```

0.0.1 Observe the training error of 0 above, and the test error of .173 above also

3. Now, prune the decision tree developed in part (1) using the data in pa2validation.txt. While selecting nodes to prune, select them in Breadth-First order, going from left to right (aka, from the Yes branches to the No branches). Write down the validation and test error after 1 and 2 rounds of pruning (that is, after you have pruned 1 and 2 nodes from the tree.)

[18]:
```
#Prune using validation v
def prune(tree, valid):
    queue = []
    queue.append([tree, valid])
    #for each node in the tree built by training set
    while len(queue)>0:
        node,valid_data = queue.pop(0)
        #if error of predicting majority label > error of predicting lable
        #replace subtree
        if mode(valid_data)[1] < calcError(node, valid_data):
            node.label = mode(valid_data)[0]
            node.decision = None
            node.thresh = None
            node.feature = None
            break
        #if node does not have label do pruning process again
        if node.label is None:
            yesSplit = valid_data[valid_data[:,node.feature]<=node.thresh]
            queue.append([node.yesBranch, yesSplit])
            noSplit = valid_data[valid_data[:,node.feature]>node.thresh]
            queue.append([node.noBranch, noSplit])
```

0.0.2 round 1 pruning

[19]:
```
prune(tree, valid)
valid_error = calcError(tree, valid)
print("validation error after one round of pruning: ", valid_error)
test_error = calcError(tree, test)
print("test error after one round of pruning: ", test_error)
```

```
validation error after one round of pruning:  0.122
test error after one round of pruning:  0.11699999999999999
```

0.0.3 round 2 pruning

[20]:

```
prune(tree, valid)
valid_error = calcError(tree, valid)
print("validation error after second round of pruning: ", valid_error)
test_error = calcError(tree, test)
print("test error after second round of pruning: ", test_error)
```

```
validation error after second round of pruning:  0.10699999999999998
test error after second round of pruning:  0.10299999999999998
```

4. Download the file pa2features.txt from the class website. This file provides a description in order of each of the features – that is, it tells you what each coordinate means. Based on the feature descriptions, what do you think is the most salient or prominent feature that predicts credit card default? (Hint: More salient features should occur higher up in the ID3 Decision tree.)

0.0.4 Feature 5, corresponding to 'PAYMENT__DELAY__SEPTEMBER' is the most salient based on it occuring higher up in the decision tree as well as being the feature for the split at the root node.

[21]:

```
features[4]
```

[21]:

```
'PAYMENT_DELAY_SEPTEMBER'
```

[ ]:

```

```