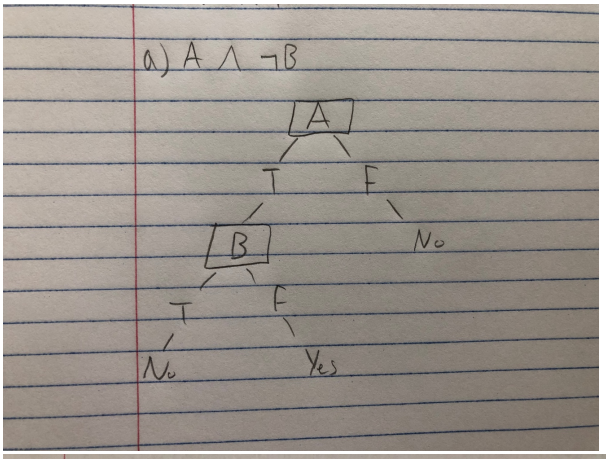
Nicholas Bonat

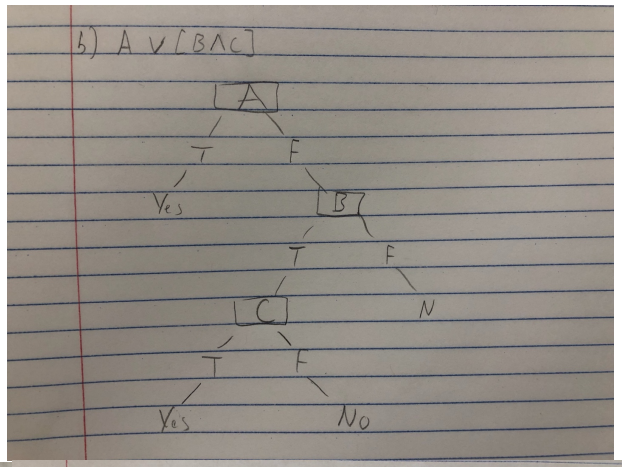
CIS 472 Machine Learning

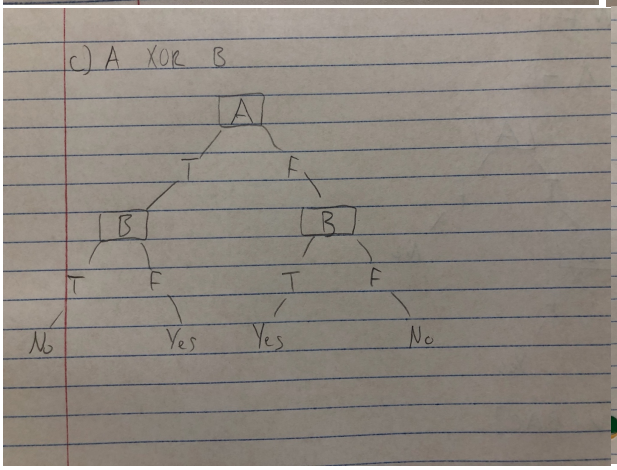
Assignment 1

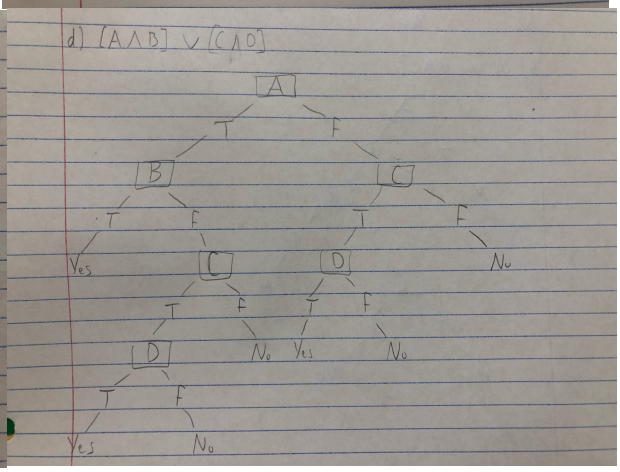
1. **Exercise 3.1**

Give decision trees to represent the following Boolean functions









2. **Consider the samples in the Play-tennis dataset from Table3.2 in Mitchell's textbook (linked above). If you calculate the information-gain for all of the attributes of this set, you will observe that the attribute "Outlook" has the largest information-gain, which is equal to 0.246. Therefore, the attribute" Outlook" is the best heuristic choice for the root node.**

a) List the labels of the new tree branches below the root node

Outlook

Sunny, Overcast, Rain

b) Partition of the data that will be assigned by ID3: Outlook {D1, D2, …, D14}

Sunny: {D1, D2, D8, D9, D11}

Overcast: {D3, D7, D12, D13}

Rain: {D4, D5, D6, D10, D14}

c) Information Gain

Sunny = 2+, 3-

-(3/5) * log(3/5) – (2/5) * log(3/5) = .97

Gain(Sunny, Humidity) = .97 – (3/5)0 – (2/5)0 = .97

Gain(Sunny, Temperature) = .97 – (2/5)0 – (2/5)1.0 – (1/5)0 = .57

Gain(Sunny, Wind) = .97 – (2/5)1 – (3/5).918 = .019

Humidity wins

Overcast = SKIP: always yes

Rain = 3+, 2-

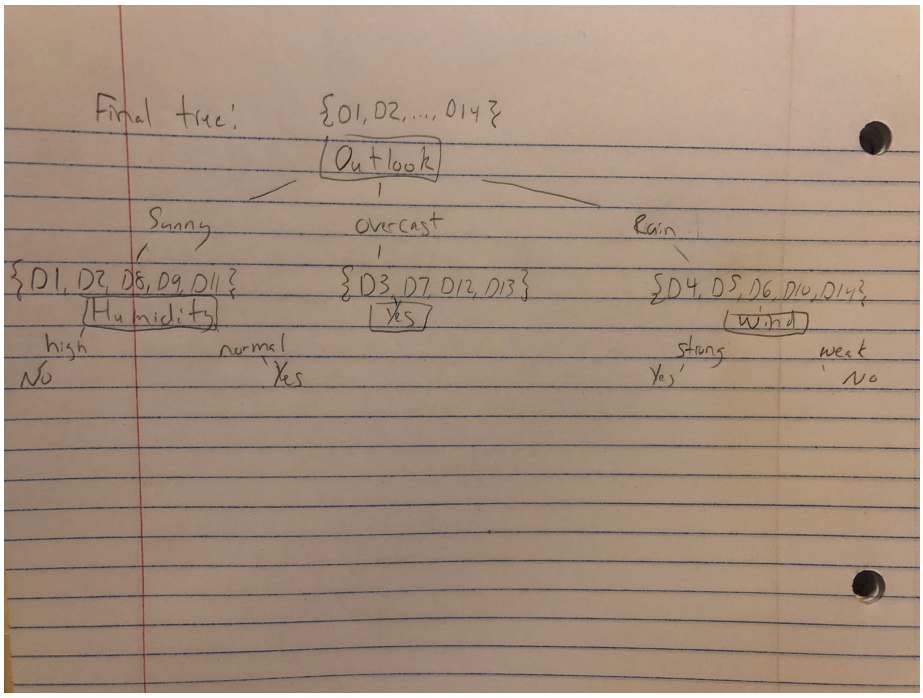
-(2/5) * log(2/5) – (3/5) * log(3/5) = .97

Gain(Rain, Humidity) = .97 – (2/5)1 – (3/5).92 = .018

Gain(Rain, Temperature) = .97 – (3/5).92 – (0/5)0 –(2/5)1= .018

Gain(Rain, Wind) = .97 – (3/5)0 – (2/5).0 = .97

Wind wins



3. **Suppose a bank makes loan decisions using two decision trees, one that uses attributes related to credit history and one that uses other demographic attributes. Each decision tree separately classifies a loan applicant as "High Risk" or "Low Risk." The bank only offers a loan when both decision trees predict "Low Risk."**

   a) In order to combine the two decision trees together, the random forest algorithm will work. To do this, several steps need to be done: select the two decision trees, select the number of samples for each tree(n), select the number of features for each tree(f), and then for each tree, select n samples with replacement from all observations, select features at random, and then train the tree using the data set of n samples with f features(Storey).

   This will give one decision tree that allows the bank to only offer a loan when both decision trees predict "Low Risk".

   b) n1 = leaves from tree1

   n2 = leaves from tree2

   The number of leaves a decision tree has is 2^h. h1 = log(n1) and h2 = log(n2).

   Upper bound: n1 * n2, this is because the worst case would be when you attach the trees end to end.

References

Storey, D. (2018, December 2). Random Forests, Decision Trees, and Ensemble Methods Explained. Retrieved from https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods