Machine Learning
COMP 5630/ COMP 6630/ COMP 6630 - D01

Instructor: Dr. Shubhra ("Santu") Karmaker
TA 1: Dongji Feng
TA 2: Souvika Sarkar
Department of Computer Science and Software Engineering
Auburn University
Fall, 2022

August 25, 2022

Assignment #1
Decision Tree

Submission Instructions

This assignment is due Tuesday, August 30, 2022, at 11:59pm. Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a PDF. Please do not include unclear scanned or photographed equations as they are difficult for us to grade.

Late Submission Policy

The late submission policy for assignments will be as follows unless otherwise specified:

1. 75% credit within 0-48 hours after the submission deadline.
2. 50% credit within 48-96 hours after the submission deadline.
3. 0% credit after 96 hours after the submission deadline.

Tasks

1. **Decision Tree Basics [50 pts]:** The goal of this assignment is to test and reinforce your understanding of Decision Tree Classifiers.

   (a) **[5 pts]** How many unique, perfect binary trees of depth 3 can be drawn if we have 5 attributes. By depth, we mean depth of the splits, not including the nodes that only contain a label. So a tree that checks just one attribute is a depth 1 tree. By perfect binary tree, we mean every node has either 0 or 2 children, and every leaf is at the same depth. Note also that a tree with the same attributes but organized at different depths

are considered "unique". Do not include trees that test the same attribute along the same path in the tree.

[5 pts] In general, for a problem with A attributes, how many unique full D depth trees can be drawn? Assume $A >> D$

(b) **[20 pts]** Consider the following dataset for this problem. Given the five attributes on the left, we want to predict if the student got an A in the course. Create 2 decision trees for this dataset. For the first, only go to depth 1. For the second go to depth 2. For all trees, use the ID3 entropy algorithm from class. For each node of the tree, show the decision, the number of positive and negative examples and show the entropy at that node.

Hint: There are a lot of calculations here. You may want to do this programatically.

Early	Finished HMK	Senior	Likes Coffee	Liked The Last Jedi	A
1	1	0	0	1	1
1	1	1	0	1	1
0	0	1	0	0	0
0	1	1	0	1	0
0	1	1	0	0	1
0	0	1	1	1	1
1	0	0	0	1	0
0	1	0	1	1	1
0	0	1	0	1	1
1	0	0	0	0	0
1	1	1	0	0	1
0	1	1	1	1	0
0	0	0	0	1	0
1	0	0	1	0	1

Table 1: Toy Data-set for Task 1: Decision Tree Basics.

(c) **[10 pts]** Make one more decision tree. Use the same procedure as in (b), but make it depth 3. Now, given these three trees, which would you prefer if you wanted to predict the grades of 10 new students who are not included in this data-set? Justify your choice.

(d) **[10 pts]** Recall the definition of the "realizable" case. "For some fixed concept class C, such as decision trees, a realizable case is one where the algorithm gets a sample consistent with some concept $c \in C$. In other words, for decision trees, a case is realizable if there is some tree that perfectly classifies the data-set.

If the number of attributes A is sufficiently large, under what condition would a dataset not be realizable for decision trees of no fixed depth? Prove that the dataset is unrealizable if and only if that condition is true.

2. **Application on Real-Word Data-set [50 pts]:** In this task, you will build a decision tree classifier using a real-word data-set called Census-Income Data Set available publicly for downloading at Dataset. This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables.

The instance weight indicates the number of people in the population that each record represents due to stratified sampling. To do real analysis and derive conclusions, this field must be used. **This attribute should *not* be used in the classifiers!!!**

One instance per line with comma delimited fields. There are 199,523 instances in the data file and 99,762 in the test file.

The data was split into train/test in approximately $\frac{2}{3}$, $\frac{1}{3}$ proportions using MineSet's MIndUtil mineset-to-mlc. Below are your tasks:

(a) **[20 pts]** Train a decision tree classifier using the data file. Feel free to use existing python/java packages/libraries you may like. You cannot touch the test file in this part. Vary the cut-off depth from 2 to 10 and report the training accuracy for each cut-off depth k. Based on your results, select an optimal k.

(b) **[15 pts]** Using the trained classifier with optimal cut-off depth k, classify the 99,762 instances from the test file and report the testing accuracy (portion of testing instances classified correctly).

(c) **[15 pts]** Do you see any over-fitting issues for this experiment? Report your observations.

Disclaimers: This assignment re-uses some materials from the publicly available website: CMU Introduction to Machine Learning Course, 10-315, Spring 2019. I personally thank Prof. Maria-Florina Balcan for sharing her teaching materials publicly. This assignment is exculively used for instructional purposes.