CS 374 Assignment #3
Decision Trees

Due the week of March 8, 2021

—

This week we explore another non-parametric technique for classifier learning: top-down induction of decision trees.

1 Decision Trees and the C4.5 Algorithm

We will focus our attention on a particular decision tree learning algorithm: C4.5, which builds decision trees by recursively selecting attributes on which to split. The criterion used for selecting an attribute is information gain.

1.1 Reading

There are many good sources of information on decision trees and the C4.5 algorithm. You might want to quickly read Alpaydin, Sections 9.1-9.3 first. Then I recommend Mitchell, Chapter 3, which should be your primary source for this topic. If you're interested in other sources, you can also look at the following:

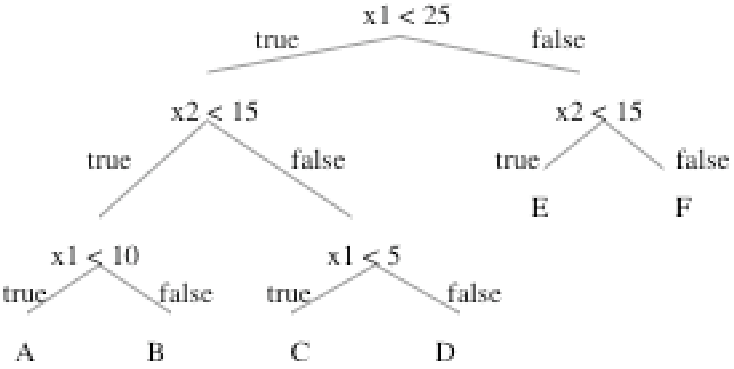
- Russell and Norvig, Section 18.3;
- Ross Quinlan's paper "Induction of Decision Trees", which appeared in Volume 1 of the journal *Machine Learning.*

1.2 Exercises, Part 1

In tutorial I will expect you to present decision trees and C4.5. You might begin by describing what a decision tree is and how it divides the attribute space into classes (for the case where attributes are real-valued). Then you might explain how C4.5 works, and demonstrate by tracing through an example. When learning a decision tree, we hope to quickly reach nodes that are "pure" in that the examples there all belong to one class. However, we often find ourselves with leaves that are not pure. How do we decide on the label to assign such a leaf? Finally, one might wonder whether successive splits of the data while building a tree can be detrimental. Explain why this is/isn't a problem.

Please prepare **written solutions** to each of the following. You might find it useful to incorporate your solutions into the tutorial presentation.

1. (From Dietterich) Consider the following decision tree:



   Draw the decision boundaries defined by this tree. Each leaf is labeled with a letter. Write this letter in the corresponding region of instance space.

2. (From an exercise by Terran Lane) Consider a two-category classification task with the following training data:

$attr_1$	$attr_2$	$attr_3$	$attr_4$	$class$
a	1	c	-1	c_1
b	0	c	-1	c_1
a	0	c	1	c_1
b	1	c	1	c_1
b	0	c	1	c_2
a	0	a	-1	c_2
a	1	a	-1	c_2
b	1	c	-1	c_2

   Construct a complete (unpruned) decision tree for this data using information gain as your splitting criterion. Please show all entropy calculations.

3. (Modified from Russell and Norvig) In the recursive construction of decision trees, it sometimes happens that a mixed set of positive and negative examples remains at a leaf node, even after all the attributes have been used.

   Suppose that you have learned a decision tree for a particular two-class problem, where 1 represents the positive class and 0 represents the negative class. Furthermore, assume that you have p positive examples and n negative examples at the leaf.

   (a) Show that the class probability $p/(p+n)$ minimizes the sum of squared errors.

   (b) Show that the solution which picks the majority class minimizes the absolute error over the set of examples at the leaf.

4. This exercise will have you consider an interesting property of the entropy function.

   For all parts of this exercise, you should assume a binary classification task, where all attributes are binary as well.

   (a) Show that the entropy function is concave. Yes, I know there are pictures of it in both Alpaydin and Mitchell. However, determining how to set up this problem in order to take derivatives will be helpful for the other parts of the exercise.

   (b) Suppose that a binary-valued attribute splits a set of examples E into subsets E_1 and E_2, and that the subsets have p_1 and p_2 positive examples and n_1 and n_2 negative examples, respectively. Show that the attribute has 0 information gain if the ratios $p_1/(p_1+n_1)$ and $p_2/(p_2+n_2)$ are the same.

   (c) A function $f(x)$ is concave on an interval [a, b] iff for any two points x_1 and x_2 in [a, b] and any λ, where $0 < \lambda < 1$,

   $$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$$

   That is, the value at the midpoint of every interval in the domain exceeds the average of its values at the ends of the interval.

   Use this to show that every attribute has non-negative information gain.

   (d) What does part c imply about the information content of the data and about the process of constructing a decision tree?

1.3 Exercises, Part 2

You won't be doing any implementation of your own this week, but you might find it useful to run C4.5 to get a sense of what it does and how well it works. To do this, you'll run Weka – open source software for data mining written in Java. (The Witten and Frank book, from which some of your readings are drawn, is a companion to an earlier version of Weka.) Weka (Version 3.6.14) is installed on the machines in the unix lab, but you can also obtain it (Version 3.8.4) to run on your own laptop. The newer version does more, but the earlier version will do everything we need. To learn more and download Weka, go to:

```
https://www.cs.waikato.ac.nz/ml/weka/
```

To run Weka's implementations of machine learning algorithms through the GUI, type the following in a terminal window when connected to one of the CS unix lab machines:

```
java -Xmx1g -jar /usr/share/java/weka.jar
```

or type the following on your own laptop:

```
java -Xmx1g -jar weka.jar
```

Click on the button that's labelled "Explorer". This will open another window that will allow you to select machine learning algorithms and datasets on which you can test them.

Begin by selecting a data set. You can do this by clicking on "Open file..." and then selecting an "arff" file of your choice. I've put some interesting data in my shared cs374 directory:

```
/home/faculty/andrea/shared/cs374/
```

You can select any file with an "arff" extension, but I would like you to begin with the nominal-valued data sets you used for the Naive Bayes implementation exercise.

Once you've selected a data file, click on "Classify" and then "Choose". In the "Trees" directory, click on J48 (which is really C4.5). Then click on "Start" at the left side of the window below "Test Options". (I had to resize the window to see the start button.) The output of the classifier will appear in the right half of the window.

Recall the analysis you did for the Titanic data set, in which you sought to determine the factors that may have contributed to a passenger's survival. Consider that question again, this time using the learned decision tree as your source for analysis. What observations can you make? **Write up** those observations, comparing them to your results from the Naive Bayes week.

If you have any trouble working with Weka or understanding how to read the output, let me know.

Please submit the answers to all of the above as a single pdf. Expect the discussion of the above to take approximately half of the tutorial session. The second half will be dedicated to discussing the "Skewing" paper, as described below.

2 Difficult Data

Of course, no learning algorithm is perfect. Your next task is to read a paper that describes a particular type of data that may be problematic for decision-tree learning.

2.1 Reading

Read "Skewing: An Efficient Alternative to Lookahead for Decision Tree Induction" by Page and Ray, which appeared in IJCAI-03.

2.2 Exercise

Write a thorough summary and critical analysis of the "Skewing" paper. As you do so, keep in mind the following questions that can serve as a guide for all reviews:

- What are the contributions of this paper to Machine Learning research?
- Do the authors back up their claims theoretically? empirically?
- Is the (theoretical or empirical) work technically sound?
- If the work is empirical, is it also reproducible?

Other important questions include:

- Discuss the relevance and importance of the contributions.
- Do the authors place the work in context? Do they cite the relevant related work?

I don't expect you to have the depth/breadth/history of knowledge in machine learning research that you'd really need to answer this second set of questions, but to the extent that you can, feel free to discuss these issues as well.

Please submit your summary and analysis as a pdf separate from the problem set above. For the tutorial session you should be prepared to present your summary and critique and to discuss it with your partner and with me.