Artificial Intelligence 2023/2024

Second assignment: Decision trees

Submission: May 17, 2024

1 Introduction

Decision Trees are a very useful way of representing knowledge in a compact way. They are also useful because their representation can be understood by non-specialists. Decision trees can be built manually, but they can also be learned from a set of observations. Decision tree models fit into the category of **supervised machine learning** algorithms, where one of the variables is well-known and classifies all observations. This variable is the one that we use to learn the model.

Using as a reference the material available in Chapter 19 of our textbook, Artificial Intelligence: a Modern Approach, by Peter Norvig and Stuart Russell (4th edition), implement an algorithm for induction of decision trees (similar to the ID3, shown in the book, Figure 19.5). Use as node selection function the **entropy** defined in page 679 (Section 19.3.3 explains how to choose "good" attributes for each level of the decision tree). It is **highly recommended** to read chapter 19 in order to understand what is a decision tree and its objectives.

This work is to be submitted via Moodle. Students must be organized in groups of 3.

2 Decision Tree Structure

The input to your program will be a set of examples in CSV (Comma Separated Value) format. This set will have several attributes (columns of your CSV table), being the last one the variable of interest for classification (Sections 19.1 to 19.3 of the textbook show an example in a tabular format). The output of your program must be in the format below, for a dataset that has 5 attributes (columns attribute1, attribute2, attribute3, attribute4, and the class variable), whose class variable has 4 different values (class1, class2, class3 and class4).

In general, `attribute` is the root of each subtree, `value#` is one of the values of the attribute (one of the branches of your tree), `class#` is the class value assigned to that branch in the tree (and corresponds to a leaf) and `counter#` is a counter of the number of examples corresponding to that tree branch. For this particular example, there are 4 attributes, where attribute1 has 3 distinct values, attribute2 has 2 distinct values, attribute3 has 2 distinct values and attribute4 has 2 distinct values. This dataset has 4 values for the class variable (class1, class2, class3 and class4). Counter# corresponds to the frequency of values of an attribute according to the class variable.

```
<attribute1>
    value1_1:
        <attribute2>
            value2_1: class1 (counter1)
            value2_2: class2 (counter2)
    value1_2: class3 (counter3)
    value1_3:
        <attribute3>
            value3_1:
               <attribute4>
                   value4_1: class4 (counter4)
                   value4_2: class2 (counter5)
            value3_2: class3 (counter6)
```

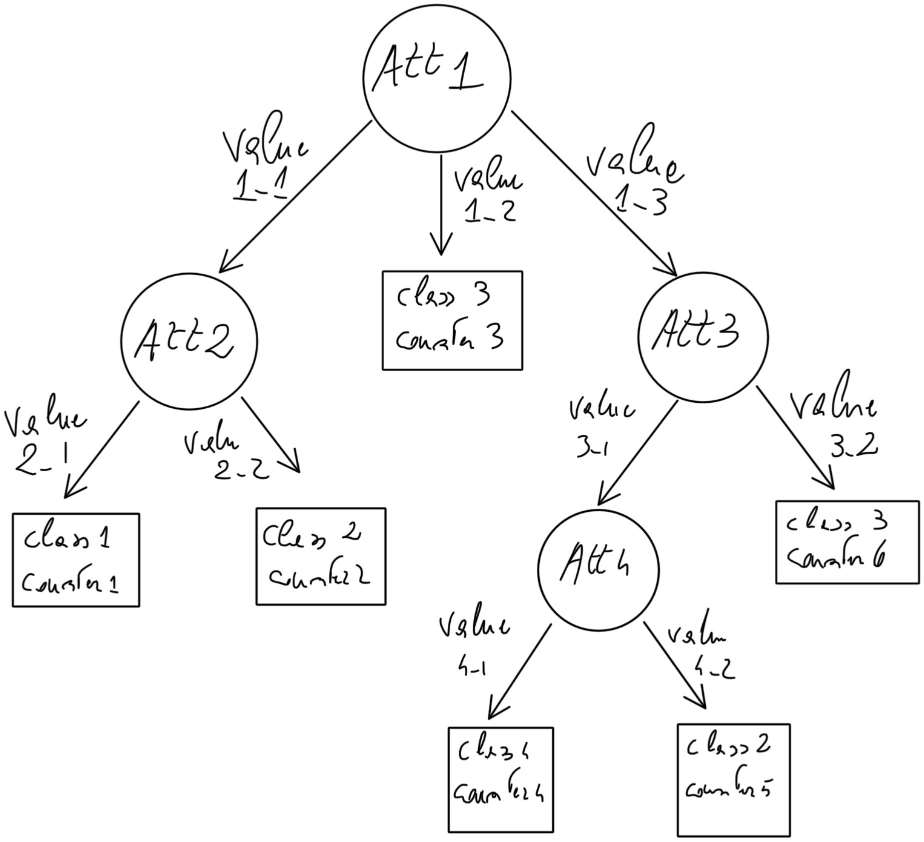

Figure 1: Decion tree

3 Datasets

In this assignment, you will be required to consider 4 different datasets for testing your decision tree implementations. All 4 datasets are available on the Moodle platform. The first three datasets are described as follows:

- `restaurant`: (example from textbook, page 675) contains information about customers and restaurants (type of food, waiting time, price etc), and the class attribute (last column) says if the customer will wait or not to eat in that restaurant. The task is to generate a decision tree (as explained in theoretical class and following the ID3 algorithm available in the textbook). This decision tree must be used later to classify (answer if the customer will wait or not) new cases.
- `weather`: contains information about climate conditions to play tennis. The task is to learn a decision tree that can decide what are the best conditions to play tennis.
- `iris`: contains **numerical** information about plants of three classes: iris setosa, iris virginica and iris versicolor. The attributes are petal length and width and sepal length and width. The task is to learn a decision tree that can tell to which class a plant belongs to, given its sepal and petal lengths and widths.

The fourth dataset that you will be required to consider is a challenge dataset that can be found in the moodle platform and in `http://archive.ics.uci.edu/ml/datasets/connect-4`, where each line corresponds to a board configuration of the connect four game. Induce a decision tree for this dataset and replace the utility function used in your Assignment 1 with the prediction of this tree to decide where to play next. You need to add a header to this file and probably an identifier to each instance, depending on your implementation). What is the impact of using the decision tree to choose the next move instead of using the utility function?

Important note: The dataset iris contains numerical values. You need to implement a way of discretizing these values in order to minimize the size of your decision tree.

4 Work to develop

The goal of this assignment is to write a program that learns a decision tree from a given training dataset using the ID3 procedure.

Your program needs to be able to read any dataset and learn the appropriate decision tree. You should **not** write one program for each one of the datasets. This means that you will need to read the CSV table from a file and store it in memory for **any** input table.

In addition to learning a decision tree, your program must also be prepared to accept as input a file with test examples, i.e., after generating your tree, you must be able to apply your tree to new examples and be able to classify them appropriately. For example, suppose you generated a tree for the restaurant problem. Now, you can enter new examples (without any class/label) and be able to give them a proper class.

Importat note: You are not be allowed to use `scikit-learn` or other libraries to automatically define and train the decision trees. External libraries will be allowed only to manage and pre-process input data.

4.1 Submission of the solution

The solution should be delivered in Moodle by May 17, 2024, at 23:59:59;

You should submit the following materials:

- Final code solution, as a notebook;
  - you should document your notebook, explaining your decisions and discussing the results obtained;
- Link for a video summary. This is a team video, but each member should participate in it. This is a very short and to-the-point video (maximum of 5 minutes), summarizing the following (you can use your notebook as background):
  - the problem;
  - your solution;
  - the results.
- Filled auto-evaluation file provided by Professors.

4.2 Presentation of the solution

Students must present their work during the practical class in the week that starts on the 20th of May, 2024 for practical classes on Thursday and 27th of May, 2024 on Monday, using the notebook (you do not need to use any additional documentation/slides).

4.3 Evaluation Criteria

Your work will be evaluated on the following criteria:

- 15% for the restaurant dataset tree implementation;
- 15% for the weather dataset tree implementation;
- 15% for the iris dataset tree implementation;
- 15% for the connect4 dataset tree implementation;
- 30% technical Skills: overall technical evaluation of the solution from a data science point-of-view;
- 10% soft-Skills: essentially - your communication skills;

4.4 Classification

This assignment represents 20% of the grade for the course (4 values). If you implement more than what is requested you can get additionally a maximum of 1 value, that can complement the grade obtained globally for the project's part.

4.5 Some Tips

Be creative in your solution! Think of how you can use certain approaches in an unusual way for example.

- Consider implementation constraints: understand the challenge well and identify any specific constraints regarding this challenge;
- Mention the constraints you are considering for the solution in the notebook;
- Work as a team: The time is very short, so we suggest that you distribute tasks well amongst the team;