Homework 9 (due ~~4/6~~ 4/13)

DS-210 @ Boston University

Spring 2022

Before you start...

Collaboration policy: You may verbally collaborate on required homework problems. However, you must write your solutions independently without showing them to other students. If you choose to collaborate on a problem, you are allowed to discuss it with at most 2 other students currently enrolled in the class.

The header of each assignment you submit must include the field "Collaborators:" with the names of the students with whom you have had discussions concerning your solutions. If you didn't collaborate with anyone, write "Collaborators: none." A failure to list collaborators may result in a credit deduction.

You may use external resources such as software documentation, textbooks, lecture notes, and videos to supplement your general understanding of the course topics. You may use references such as books and online resources for well known facts. However, you must always cite the source.

You may **not** look up answers to a homework assignment in the published literature or on the web. You may **not** share written work with anyone else.

Submitting: Solutions should be submitted via Gradescope. The entry code is 3Y85PZ.

Grading: Whenever we ask for a solution, you may receive partial credit if your solution is not sufficiently efficient or close to optimal. For instance, if we ask you to solve a specific problem that has a polynomial–time algorithm that is easy to implement, but the solution you provide is exponentially slower, you are likely to receive partial credit.

Late submission policy: No extensions, except for extraordinary circumstances. We accept submissions submitted up to one day late, but we may deduct 10% of points.

Questions

To solve problems in this homework, you should use Rust. Your solution to the homework should consist of

- a compilable Rust source file (`.rs`) solving Question 1,
- a report (the pdf format is recommended) that answers sub-questions denoted by "**Report**" below.

1. **(40 points)**

   Input: Your program should read a set of points in $\mathbb{R}$ with labels from `data.txt`. Each line of `data.txt` describes one point. More specifically, the i-th line consists of two numbers x_i and z_i, where x_i is an integer s.t. $|x_i| \leq 100,000,000$ and $z_i \in \{0, 1\}$. x_i is the coordinate of the point and z_i is its label.

   Your task: Write a program that reads the data and determines a decision tree with **at most two leaves** that performs best at predicting z_i based on x_i on this set of points. The program should output the decision tree and its accuracy on the input data set.

   Report: Explain why your solution works and what complexity it has as a function of the number of points (denote this quantity by n).

   Sample input and output: For the following `data.txt`:

   ```
   -15 0
   15 1
   -5 1
   5 0
   ```

   the following output is a correct solution:

   ```
   if x >= 10
       output 1
   else
       output 0

   accuracy: 0.75
   ```

2. **(Optional, no credit)**

   Report: How much time did you spend on this homework? The answer will have no impact on the credit you receive, but it may help us adjust the difficulty of future homework assignments.

3. **(Optional, no credit)**

   Report and/or code: How can you solve Question 1 when more than two leaves are allowed? What is the complexity of your solution as a function of the numbers of leaves and points?