SCC461 – Programming for Data Scientists

Leandro Marcolino

Week 8

Assignment

Deadline: Monday, 04/12, 9am
Upload on Moodle your code, your test cases (with the output), and your short reflection.

1. Decision Tree (4%)

   In this assignment, you must implement and test a Decision Tree class. Given a dataset of items and labels, we can learn a Decision Tree in order to perform classification. We will consider a binary tree, where at each node we must find the best partition that minimizes the Gini index:

   $$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2),$$

   where D_1 is the subset on the left son of the node, and D_2 is the subset on the right son of the node. $Gini(D_1)$ and $Gini(D_2)$ are calculated as follows:

   $$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

   where:

   $$p_i = \frac{|C_{i,D}|}{|D|} = \frac{\text{\# of items with class } C_i}{\text{total \# of items in } D}$$

   For this assignment we will consider only two possible features: the first is a continuous number between 0 and 10, the second is a categorial variable with the following possible values: 0, 1, 2. As usual, you should test all midpoints of continuous variables and all subsets of categorial variables. Only the test $x_i \leq y$ is used for continuous variables splits, and $x_i \in S'$ for categorial variables. We will also consider only two possible labels: 0 or 1.

   For more details, please refer to the slides in Week 7 of SCC403, and the book "Data Mining: Concepts and Techniques – Chapter 8.2".

2. Inheritance (1%)

Re-write your FibonacciQueue and your PriorityQueue from Week 7 assignment using inheritance. That is, you should have a Queue class (which can be the one shared in the class), and inherit from that class in order to create a FibonacciQueue and a PriorityQueue.

As mentioned in class, you must write a short text reflecting how you approached these problems. You must also report who you discussed with, what you searched online, who you helped, etc. Discussions are allowed, and looking for online materials, books, etc, is allowed. However, directly copying full Python code is not allowed.