Homework 6

2-AIN-150, Winter 2018

Deadline: 3.12.2018, 23:59, moodle.uniba.sk/fmfi

Before you start solving the homework, please read the general instruction at the end of the document. Submitted solutions should be your own. Do not copy and do not try to find solution in literature or over the internet.

Finding a good split

The task is simple. We are doing regression, we have one input attribute and one output. Write a function, which fits decision tree of depth one over the dataset. Also try to make this function as efficient as possible.

In other words: Given pairs: $(x_1, y_1), \ldots, (x_n, y_n)$. For given s we define sets $A_- = \{y_i | x_i < s\}$, $A_+ = \{y_i | x_i \geq s\}$ (A_- has those y_i, which has corresponding x_i less than s). Find such s, for which A_- and A_+ are nonempty and the value: $\frac{\texttt{Var}(A_-)|A_-| + \texttt{Var}(A_+)|A_+|}{n}$ is smallest as possible. ($\texttt{Var}(X) = \frac{1}{|X|}\sum_{x \in X}\left(x - \frac{1}{|X|}\sum_{y \in X} y\right)^2$, aka variance)

There are two datasets in the package. One small, one larger. Your program should be fast on both of them.

Your program can use any libraries for basic math, matrix operation, matrix inversion, solving systems of linear equations, and calcualating numerical or symbolic derivatives. You are forbidden to use scikit-learn library and its equivalents in other languages.

Python instructions There is `template.py` in package. You should fill out the function `get_s(x, y)`. Program can be runned using `python template.py <input file>`.

Instructions for other languages Input files contain number n and then n lines: on each line one pair x_i, y_i. Please provide clean output from your program.

General instructions

You should submit homework via email with subject listed in title. Add your code to the email attachement.

Ideally submit your homeworks in Python (fill out required functionality from assignment). You can use different language if you really want, but you need also to add auxiliary functionality like reading input and output. But your solution should be runnable under Linux using open source software.