Finding a good split

3. júla 2016

We are training a regression on data with one attribute. Write a function, which finds the best possible split of the data (same thing which decision tree would do in its root). Try to make the function as efficient as possible.

More specifically: You are given pairs:

$$(x_1, y_1), \ldots, (x_n, y_n)$$

.

For given s we define sets:

$$A_- = \{y_i | x_i < s\}$$

$$A_+ = \{y_i | x_i \geq s\}$$

(those y_i where $x_i < s$ end up A_- and vice versa).
Find s, for which A_i and A_+ are non empty and value is smallest as possible:

$$\frac{\mathtt{Var}(A_-)|A_-| + \mathtt{Var}(A_+)|A_+|}{n}$$

($\mathtt{Var}(X) = \frac{1}{|X|} \sum_{x \in X} \left(x - \frac{1}{|X|} \sum_{y \in X} y \right)^2$)

You are given inputs and template of the solution in Python (you can also you Java or Scala).