RWTH Aachen
Lehrstuhl für Informatik 9
Prof. Dr. T. Seidl

January 18th, 2016

Exercise 9 for the lecture
Data Mining Algorithms
WS 2015/2016

Hand in your solutions on January 25th before the lecture. The tutorial for this exercise will be held on January 29th.

Exercise 10.1) Linear Regression — 5 points

Consider a set of N observations of the form (x, y), with inputs $x \in \mathbb{R}$ and outputs $y \in \mathbb{R}$, and consider the linear function $f: \mathbb{R} \to \mathbb{R}$, $f(x) = \beta_0 + \beta_1 x$, which estimates the output ($f(x) = \hat{y}$). Show that the sum of squared errors

$$g(\beta) = \sum_{i=1}^{N} \left(y_i - f(\boldsymbol{x_i})\right)^2$$

is minimized by the coefficients β_0 and β_1 as follows:

$$\beta_1 = \frac{cov(x,y)}{var(x)} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x},$$

where $\bar{x}, \bar{y}$ denote the means as usual.

Note: Since the SSE is a convex function, you only need to set the first derivative to zero and solve the resulting equation.

Exercise 10.2) Regression Trees — 8=4+4 points

Consider the following data samples of the form (x, y), where the input value is $x \in \mathbb{R}$ and the output value is $y \in \mathbb{R}$:

P1=(2,13), P2=(3,14), P3=(4,14), P4=(5,12), P5=(6,10)

Search for the first best split, by computing all possible splits, and then decide whether the split is significant or not by using the impurity ratio, with $\tau_0 = 0.5$.

a) Use as impurity measure the standard deviation of the output.

b) Use as impurity measure the variance of the residuals.

Exercise 10.3) Hinging Hyperplane Models **8 points**

Consider the following data samples of the form (x, y), where the input value is $x \in \mathbb{R}$ and the output value is $y \in \mathbb{R}$:

P1=(2,4), P2=(3,5), P3=(4,6), P4=(5,7), P5=(6,8), P6=(7,8), P7=(8,8), P8=(9,8), P9=(10,8), P10=(11,8), P11=(12,6), P12=(13,4), P13=(14,2).

Iteratively fit a hinge, starting with an initial partitioning in the input space {P1,P2} and {P3,P4,…,P13}. For each iteration: write down the two resulting partitions, their corresponding linear models and the computation of the new hinge. At the end, write down the resulting piecewise linear function.