CS 4780/5780 Homework 8

Due: Thursday 11/29/18 11:55pm on Gradescope

Problem 1: Regression Trees

(a) You are given a dataset $D = \{(-3,-20),(-2,-20),(-1,-17),(0,15),(1,25),(2,26)\}$ and you want to build a regression tree for this dataset. Recall that the impurity for the regression tree model is defined as

$$L(S) = \frac{1}{|S|} \sum_{(x_i,y_i)\in S} (y_i - \bar{y}_S)^2,$$

where $\bar{y}_S = \frac{1}{|S|}\sum_{(x_i,y_i)\in S} y_i$. Draw the regression tree T_0 built by the ID3-Algorithm which was introduced in class.(There are multiple correct thresholds. Choose one of them to draw.)

(b) We keep the definition of $L(S)$ as (a). Prove that $L(S) \geq \frac{|S_1|}{|S|}L(S_1) + \frac{|S_2|}{|S|}L(S_2)$, where $S_1 \cup S_2 = S$ and $S_1 \cap S_2 = \emptyset$. This conclusion tells us the impurity of a regression tree never increases after one split.

Problem 2: Normalization Update in Adaboost

In the Adaboost, we keep $\sum_{i=1}^n w_t^i = 1$. In the iteration t of the algorithm, we update w_t^i as follow:

$$w_{t+1}^i \leftarrow \frac{w_t^i \cdot e^{-\alpha_{t+1}h_{t+1}(x_i)y_i}}{2\sqrt{\epsilon_{t+1}(1-\epsilon_{t+1})}}$$

where $\alpha_{t+1} = \frac{1}{2}\log\left(\frac{1-\epsilon_{t+1}}{\epsilon_{t+1}}\right)$ and $\epsilon_{t+1} = \sum_{i:h_{t+1}(x_i)\neq y_i} w_t^i$. Prove that if $\sum_{i=1}^n w_t^i = 1$, $\sum_{i=1}^n w_{t+1}^i = 1$, i.e. $\sum_{i=1}^n w_t^i \cdot e^{-\alpha_{t+1}h_{t+1}(x_i)y_i} = 2\sqrt{\epsilon_{t+1}(1-\epsilon_{t+1})}$. (Remember in the Adaboost, $h_{t+1}(x_i), y_i \in \{+1,-1\}$.)