Decision Trees

1. Task

We are given a univariate regression data set, which contains 272 data points about the duration of the eruption and waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We cant to use decision trees to do regression on this data.

2. Implementation

We implement decision tree regression, where a split in each node is decided by the following algorithm: Let $X = \{x_1, x_2, ..., x_N\}$ be the data reaching the current node, and $Y = \{y_1, y_2, ..., y_N\}$ be their target values respectively. We generate $N - 1$ split candidates $S = \{s_1, s_2, ..., s_{N-1}\}$ where $s_i = (x_i + x_{i+1})/2$. We then calculate the split errors and choose the minimum. To calculate the error, we split X into X_L and X_R where $X = X_L \cup X_R$ and $Y = Y_L \cup Y_R$ respectively. Let a mean of set S be shown as S'. Then our error function is:

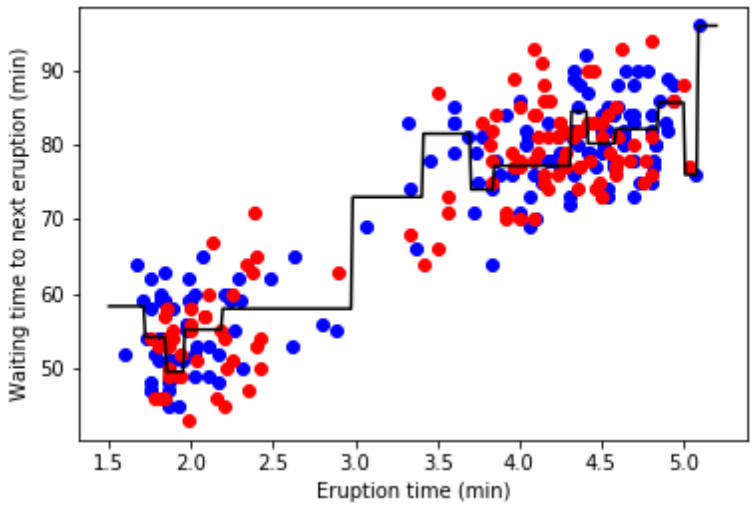
$$\frac{\sum_{i=1}^{|Y_L|}(Y_{Li} - Y'_L)^2 + \sum_{i=1}^{|Y_R|}(Y_{Ri} - Y'_R)^2}{N}$$

Notice that $N = |Y_L| + |Y_R|$. We choose the split with minimum error, and recursively construct the nodes. We make a node terminal when one of the splits are empty, which means a split is not required. When we have $N < P$ where P is the pre-pruning parameter, we let that node to be a terminal node, i.e.: a leaf. We also make a node terminal if $N = 1$ or $N > 1$ but all data is same, so in other words there is only one unique data.

Our results are on the next page.

Here is our result with $P = 25$:

FIGURE 1. Results with $P = 25$



We also calculate the RMSE for several pre-pruning parameters, and here is our observation:

FIGURE 2. RMSE versus Pre-pruning Parameter