Homework 2

September 15, 2018

1 Homework 2

Question 1 *Best K-division*

Let N be a node of a regression tree, design a greedy algorithm to find the best division of the data set, which ensures the greatest target function loss.

The target function is defined as:

$$\sum_{j=1}^{T}(\frac{-2B_j^2}{2A_j+\lambda})+\gamma T+Const$$

Where I_i represents a node, y_i is the real value of a record, T is the number of leaf nodes generated, $A_j=|I_j|$, $B_j=\sum_j y_j$.

We assume that the number of elements of the node is $NodeSize$.

Algorithm 1 *Faster Enumeration*

Firstly we fix the sequence of all the elements.

A basic idea is to enumerate all posible ways of division. By dynamic programming, we can lower the cost.

We use 2-d array Result to record intermediate results. Result[j][i] means the lowest target function value from division(i,j), which means the last division happend before the gap between element 'i' and 'i+1' with j divisions in total.

Thus we can get the algorithm represented by pseudocode.

Algorithm 1 Faster Enumeration

```
//Prepare Data
Let Result[<0][]=INF;
Let Result[][<0]=INF;
Let Result[>=NodeSize][]=INF;
Let Result[][>=NodeSize]=INF;

//Dynamic Programming
for(j=0;j<NodeSize;j++)
    for(i=0;i<NodeSize;i++)
    {
        Result[j][i]=
            min{Result[j][i-1],(Result[j-1][i]+Loss(j,i))};
        Update recordofdivisions[j][i];//record the position of divisions.
    };

//Pick Result
Return min(Result), as well as the way of division;
```

'Loss(j,i)' is the loss gain from adding a division at the ith gap. It is related to 'recordofdivisions[j-1][i]'.