Stat365/665 (Spring 2015) Data Mining and Machine Learning **Lecture: 16**

Data Mining and Machine Learning

Lecturer: Sahand Negahban *Scribe: Leon Lixing Yu*

1 Boosting and overfitting

The decision tree I used is matlab function $fitctree$, that takes inputs, Y, X, $weights$, and $maxNumSplits$. The maximum number of splits are used to control the depth of the tree.
If the $maxNumSplits$ is 3, the depth of the tree is 2. likewise, if the $maxNumSplits$ is 4, the depth of the tree is 3. Since we are given the depth of the tree as 3, I choose the $maxNumSplits = 4$.
The code for this problem is attached in $Appendix - A$.The figures below shows the test errors while changing the number of iterations with fixed tree depth.

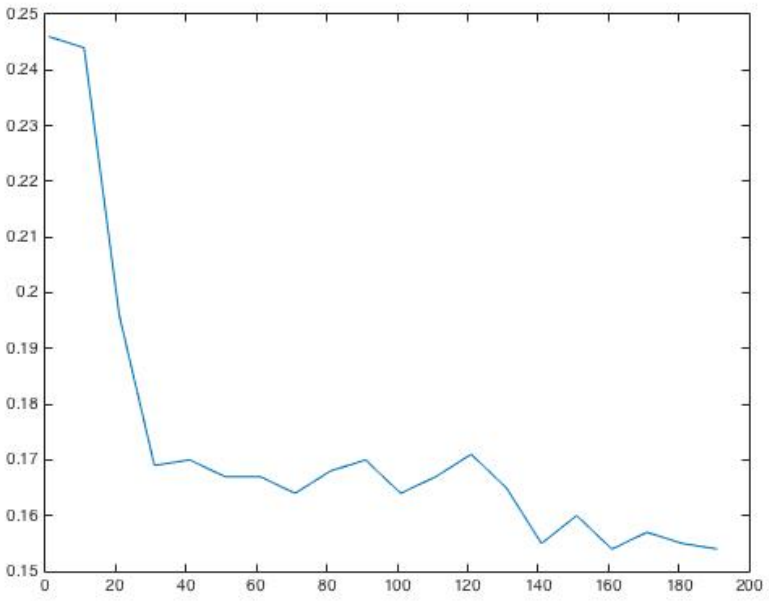

Figure 1: test error for upto 40 iterations.I can see a dip at 11th iteration. The testing error and then fluctuates alone the similiar domain.To understand why, I plotted training error to see the convergence.

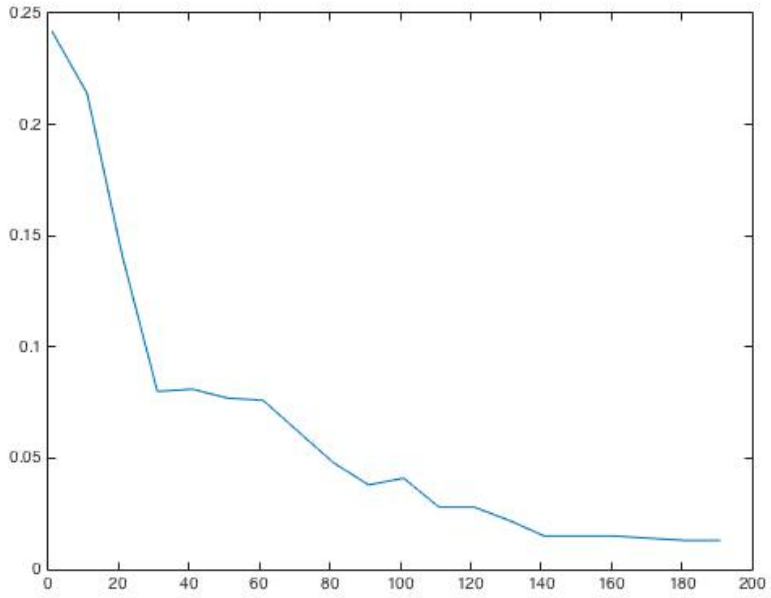

Figure 2: training error for upto 40 iterations.I can see at around 20th iteration, the boosting converges to 0 error, meaning everything after 20th iteration is for sure overfitting. Also, since the dip happens at 9th iteration during testing, I claim that the best fit is around 9th iteration, and any work done beyond 9th iteration is considered overfitting.

The figure below shows the graph for testing error with incremental tree depth while fixing the iteration number to 3

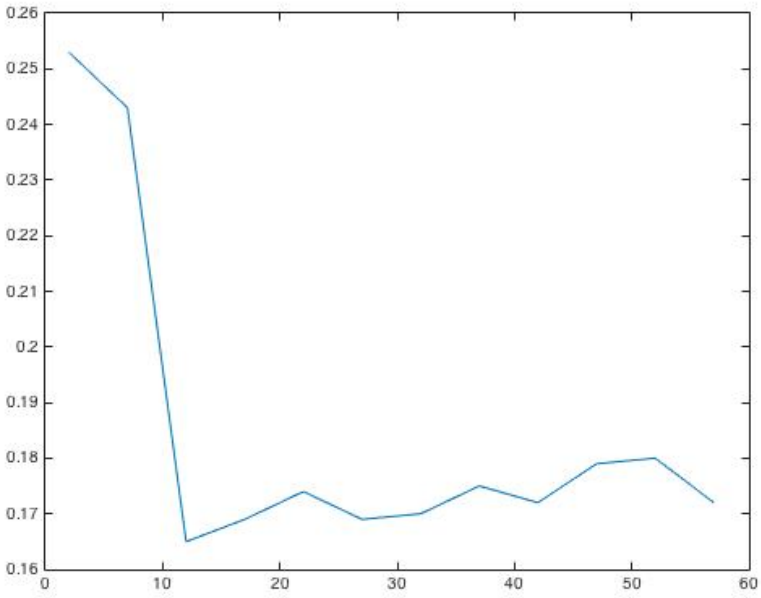

Figure 3: Training error for upto 60 splits. The iteration is fixed to 3. I see it reachs the best testing error (17% ish) while number of splits equals to 10. If i increase the depth even further, it is the case of overfitting as I do not see any performance gain; instead, the testing error was increased to 18%

2 boosting for regression

The algorithm is shown below:

m is the number of iteration for $t = 0 : m$