Statistical Machine Learning (STAT W4400)

Homework 2

Linan Qiu
`lq2137`

October 27, 2015

As usual, code is available at https://github.com/linanqiu/stat-w4400-homework

1 Adaboost

1.1 Implement Adaboost in R

Done in `adaboost.R`

1.2 Implement decision stump `train` and `classify`

Done in `stump.R`

To generate weak learners, I guessed a θ, then if the θ produces a cost greater than 5, I took the negative m.

1.3 Run Algorithm on USPS Data

Done in `adaboost.R` with K crossfold validation (defaults to 5).

1.4 Plot Training and Testing Error

Plots shown next page.

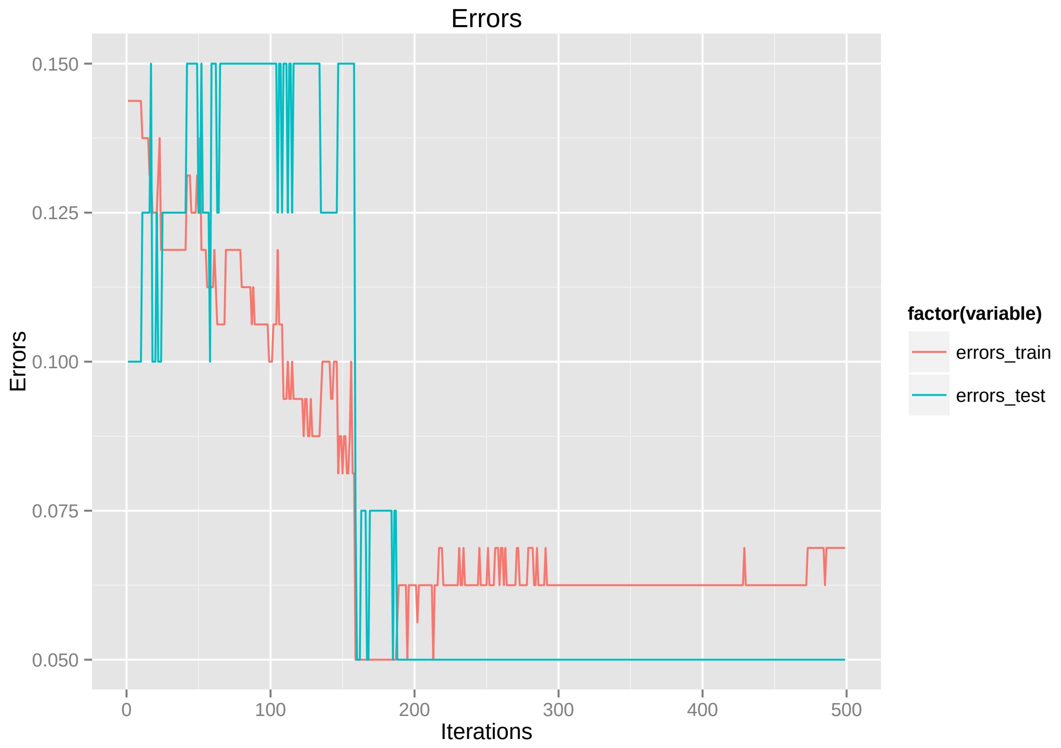

Figure 1: Except for a weird spike in test error after around 100 iterations (probably due to the weak nature of the learners) we see a rather fast decrease in test error.

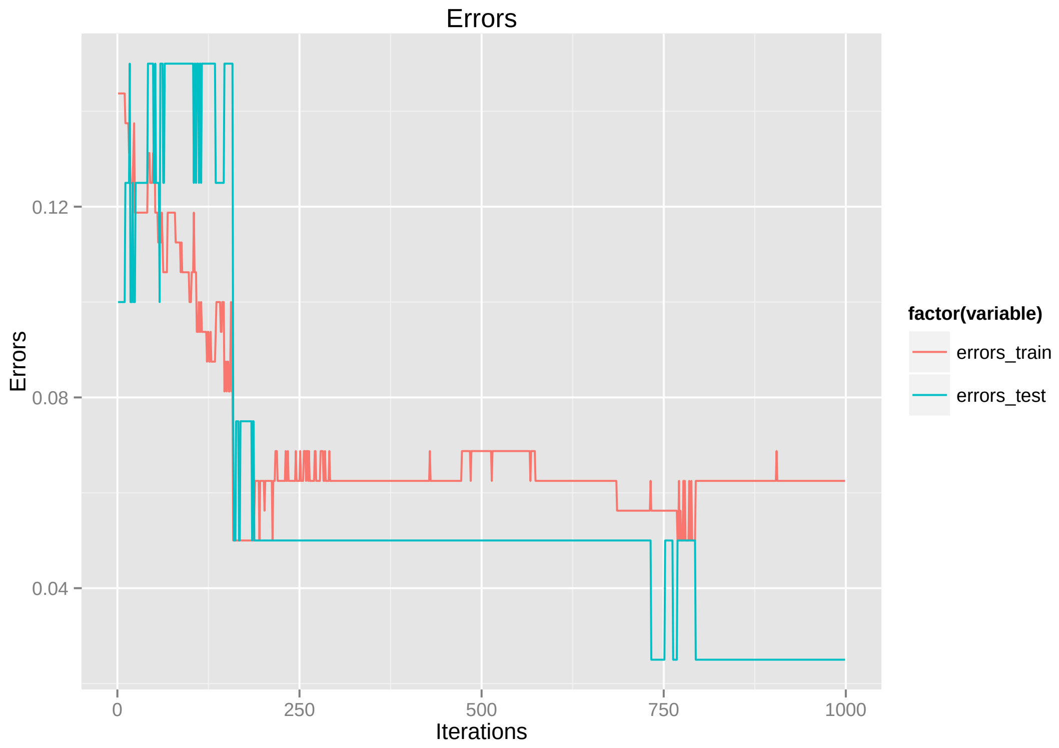

Figure 2: The decrease continues into 1000. However, the quantized nature of the jumps (due to the small data size) is not helpful at all.

2

2.1

The left cost function encourages sparse estimates. It prefers points in $\hat{\beta}$ that have either β_1 or β_2 but not both, whereas the one on the right tends to encourage the opposite. This is already evident in the illustration, where x_3 and x_5 intersects with $\hat{\beta}$.

2.2

- For $q = 0.5$, x_3 minimizes the cost as it is the only point to intersect the constraint region.
- For $q = 4$, x_3 and x_5 intersects, but x_4 minimizes the cost since it lies within the constraint region, and would be lower cost than x_3 and x_5 that satisfies the constraint but lie on the border.