Big Data and Analytics Programming: assignment 2

Toon Nolten
toon.nolten@student.kuleuven.be

February 6, 2016

1 Bugs

My implementation of very fast decision trees isn't faster than that of the logistic regression. This leads me to believe it is somehow flawed but as I do not have reference data on how much faster (or not) it should be, I have not investigated this further.

I believe the resolution for the learning curves isn't high enough to be very interesting. I tried a linear sequence of stream sizes instead of the exponential and some interesting features showed up. However we were asked not to change the code of the methods that were already implemented, so I have reverted this change.

There was one bug in the main method of the logistic regression class: the output file had the extensions ".nb.acc" instead of ".lr.acc". I chose to alter the code in this instance, complying with the assignment.

2 Learning curves

The learning curves for logistic regression and very fast decision trees are show in figures 1 and 2 respectively. What's remarkable is that VFDT gains a lot of accuracy early on (10 to 100 examples), while logistic regression only improves significantly after many (10000 to 100000) examples.

As you can see the learning curve for Naïve Bayes is missing. My implementation of Naïve Bayes was faulty and I decided to drop it.

The learning curves were plotted on the big data set without noise. I disregard noise because I'm comparing the power of the methods relative to one another, not their practical applicability and while the figures may seem more interesting with the noise, it becomes hard to explain why.

3 Experiment

3.1 Questions

The most important parameter for logistic regression seems to be the learning rate. How does the learning rate affect the learning curve? For VFDT the most important parameter seems to be δ. How does δ influence the learning curve?

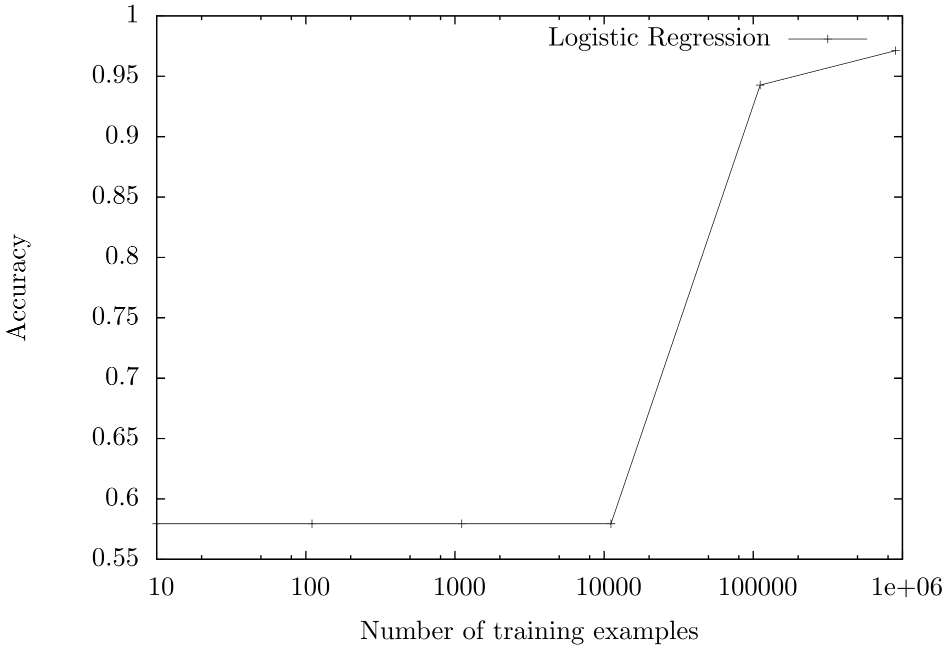

Figure 1: Learning Curve for Logistic Regression

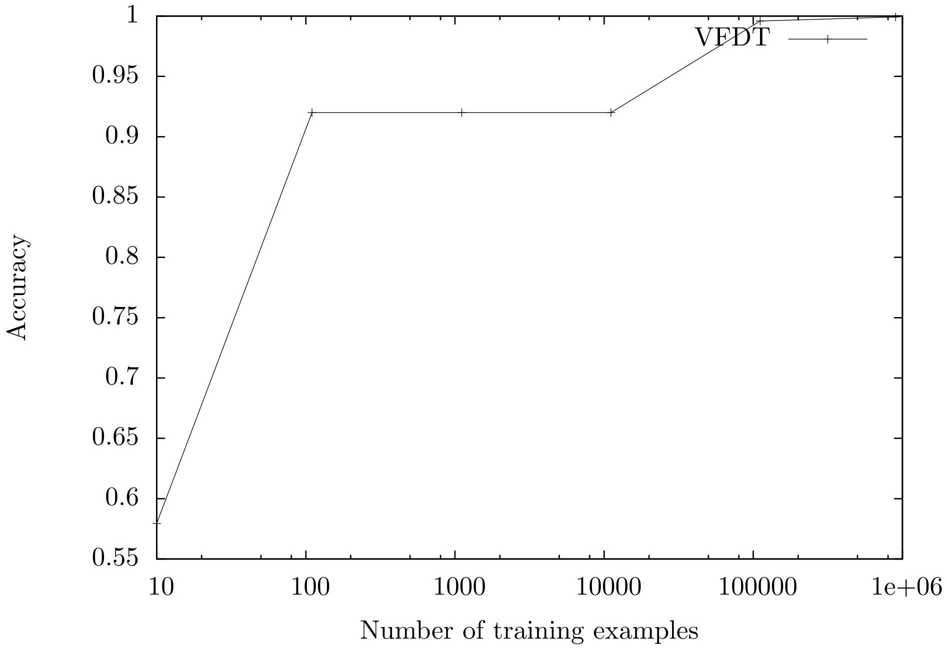

Figure 2: Learning Curve for VFDT

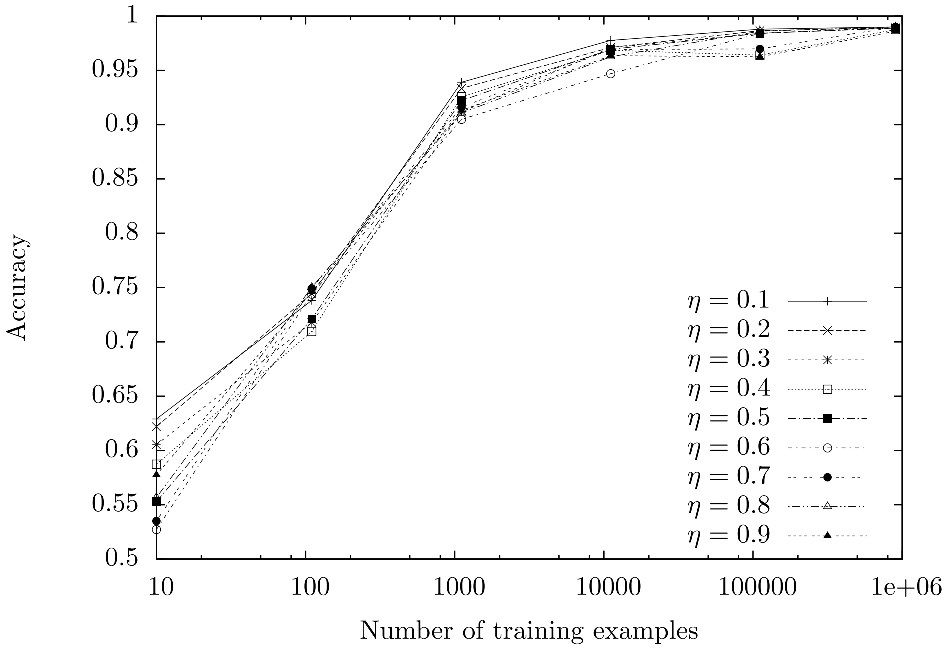

Figure 3: Learning Curves for Logistic Regression

3.2 Setup

For values of the learning rate η, respectively vfdt's δ, uniformly distributed over the open interval $]0, 1[$, run the appropriate algorithm and plot the learning curve. All other parameters remain as before.

3.3 Results

The results are quite boring. For logistic regression lower learning rates show better accuracy with fewer examples but past about 100 there doesn't seem to be much difference, see figure 3. This is opposite of what you'd expect from a learning rate. However the learning rate used to plot the curve in figure 1 is much lower and shows the expected corresponding delay in learning.

In figure 4 there's even less to see. They are all very similar. The one that stands out, $\delta = 0.1$, and the one in figure 2 do show the expected effect. If δ becomes very small, requiring a high probability that a split is correct, it becomes harder to find a good split. This delays the increase in accuracy by making a good split. This homogeneity is probably due to the data set. The lack of noise means that splits that look good at a certain level of confidence also look good at higher levels of confidence.

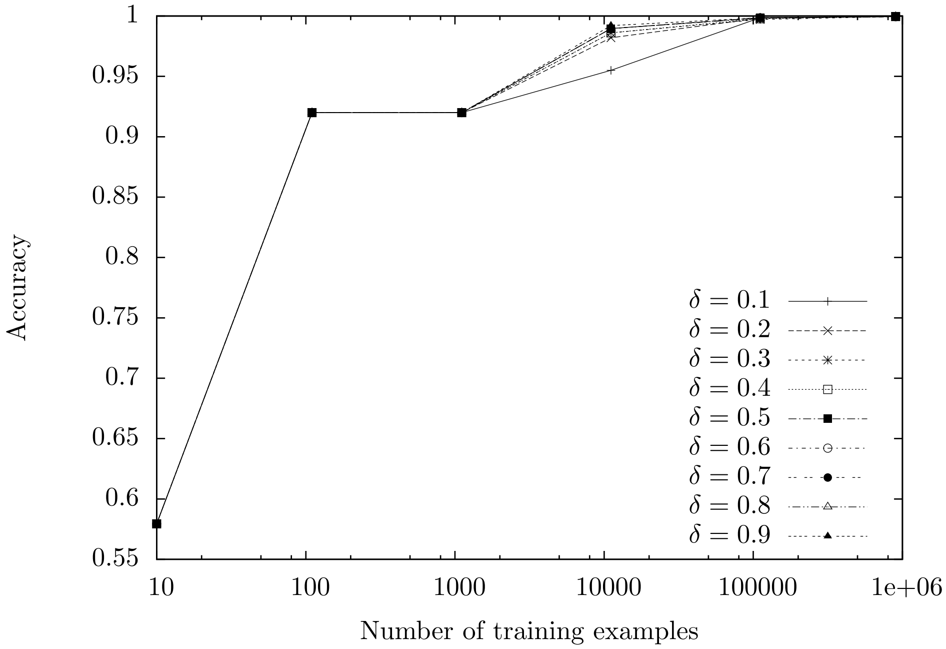

Figure 4: Learning Curves for VFDT