# Theoretical questions

We look on the following training set:

We look on the first iteration of ID3 with the information gain:

Therefore .

We want to calculate which split the algorithm will choose,

To do that we calculate

for .

Let us calculate the pre-split error:

:

This means that algorithm will choose to split by .

This means that during the classification

will go to the same branch.

If the algorithm will choose to split by then and will go to the same branch.

If the algorithm will choose to further classify by then and will go to the same branch.

Because we stop the algorithm in the next iteration this means we will get false classification with at least one sample, this means that the training error of the resulting decision tree is at least 1/4.

Let us look at the following tree:

Where at each node we ask if , the left branch marks false result, and the right branch marks positive results.

We can see that this tree is with depth 2, and that on the training set we get that:

On the result is 1.

On the result is 1.

On the result is 0.

On the result is 1.

Thus, this tree achieves zero training error.

Let there be distributions over .

Let us consider the concave function .

From Jensen’s inequality, we obtain that

Therefore, .

Let there be distributions over s.t independent from and independent from

Where we used that .

Let us look on the iteration of AdaBoost.

We assume the .

Let us calculate the error the by the distribution :

Where we used that , and when .

* 1. S

We assume by contradiction that .

Using article a., we notice that .

But this contradicts our assumptions on the weakly learner.

Let be the training set, a hypothesis class.

We assume that there is , hypotheses and

coefficients s.t for which the following holds:

For all .

Let there be distribution over .

We take the expectation of both sides:

Because , this means that there is s.t

We now compute the expectation w.r.t of

This means that

* 1. D

Let there be training set that is realized by a d dimensional rectangle .

Using the hint, we set .

We will compose hypotheses:

For we define:

And we use hypotheses that is constant -1, i.e., .

Totally we have hypotheses from .

Let there be positive sample in :

We know that for all

Therefore,

Let there be negative sample in .

This means that is not in the d dimensional rectangle.

This means that there is s.t , this means that

.

If then

And if then

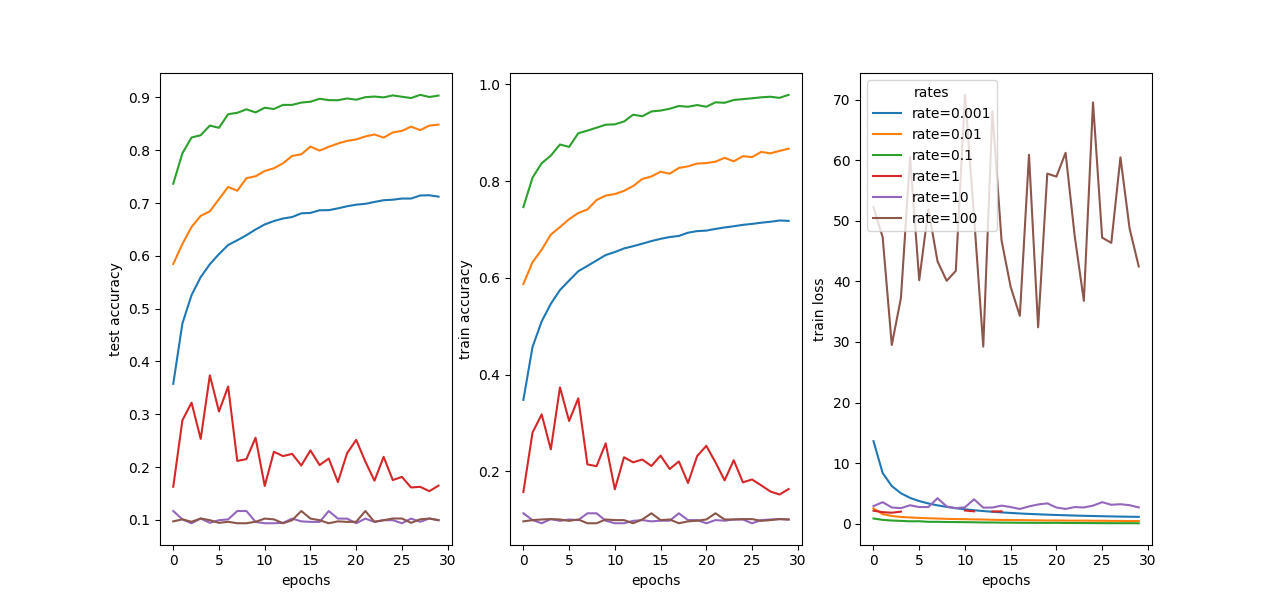
This means that

Because we left only with literals .

This means that

So, if we denote we get what we wanted.

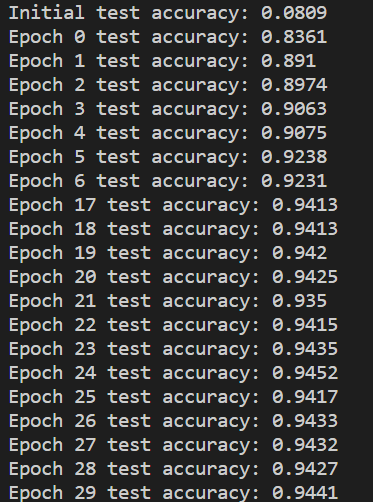
# Programming questions:

1. 

We can see that in some cases our learning was good, this happened when the learning rate was relatively small (smaller than ).

When the learning rate was large, we can see that we haven’t converged to an optimum, perhaps this happened because the leaning step in SGD was too large and caused us to miss a critical point.

When the learning rate was too small, we haven’t managed to converge to an optimum, but it looks like with more iterations SGD will finally converge with this learning rate.

1. Using the entire dataset, 30 iterations, learning rate of 0.1, batch size 10, and architecture of [784, 40, 10] the test accuracy in the final epoch is .