Convergence of Adalines

We can show convergence of adalines similar to perceptrons.

The adaline learning rule converges to L_2 loss, approximately.

We can show that the distance between $\mathbf{w}$ and the optimal $\mathbf{w}^*$ decreases when $\mathbf{w}$ makes a mistake.

Let the identity function $o = u$ be the activation function:

Let the learning rule be:

$\mathbf{w} \leftarrow \mathbf{w} + \eta(d - o)\mathbf{x}$

Let the error function be:

$$E(d, o) = (d - o)^2 = (d - u)^2$$

Let $u = \mathbf{w} \cdot \mathbf{x}$ and $u^* = \mathbf{w}^* \cdot \mathbf{x}$.

Distance to Optimal Weights

Measure the distance between $\mathbf{w}$ and $\mathbf{w}^*$ by:

$$\|\mathbf{w} - \mathbf{w}^*\|^2 = \sum_i (w_i - w_i^*)^2$$

After an example, $\mathbf{w}$ is updated to $\mathbf{w}' = \mathbf{w} + \eta(d - u)\mathbf{x}$.

We want to know when $\mathbf{w}'$ is closer to $\mathbf{w}^*$, so we analyze

$$\|\mathbf{w} - \mathbf{w}^*\|^2 - \|\mathbf{w}' - \mathbf{w}^*\|^2$$

and determine when it is positive.

$$\begin{aligned}
&\|\mathbf{w} - \mathbf{w}^*\|^2 - \|\mathbf{w}' - \mathbf{w}^*\|^2 \\
&= \sum_i \left((w_i - w_i^*)^2 - (w_i' - w_i^*)^2 \right) \\
&= \sum_i \Big((w_i - w_i^*)^2 \\
&\qquad - (w_i + \eta(d - u)x_i - w_i^*)^2 \Big) \\
&= (d - u) \sum_i \left(2\eta x_i (w_i^* - w_i) - (\eta x_i)^2 \right) \\
&= 2\eta(d - u) \left(\sum_i w_i^* x_i - \sum_i w_i x_i \right) \\
&\qquad - \eta^2 (d - u)^2 \sum_i x_i^2 \\
&= 2\eta(d - u)(u^* - u) - \eta^2 (d - u)^2 \|\mathbf{x}\|^2
\end{aligned}$$

By rearranging terms and using $X \geq \|\mathbf{x}\|$, the last expression is greater than or equal to:

$$\eta(2 - \eta X^2)(d-u)^2 - 2\eta(d-u^*)^2 \\ - 2\eta(d-u^*)(u^*-u)$$

This is greater than or equal to:

$$\eta(2 - \eta X^2)(d-u)^2 - 2\eta(d-u^*)^2 \\ - \eta(d-u)^2/2$$

which is equal to:

$$\eta(3/2 - \eta X^2)(d-u)^2 - 2\eta(d-u^*)^2$$

Interpretation of Result

The progress towards $\mathbf{w}^*$ is positive if:

$$(d-u)^2 \geq \frac{4}{3 - 2\eta X^2}(d-u^*)^2$$

Note $(d-u)^2$ is $\mathbf{w}$'s error, while $(d-u^*)^2$ is $\mathbf{w}^*$'s error.

If we choose η small enough, say $1/(2X^2)$ or less, then $\mathbf{w}$'s error will converge twice optimal or less.

Experiments

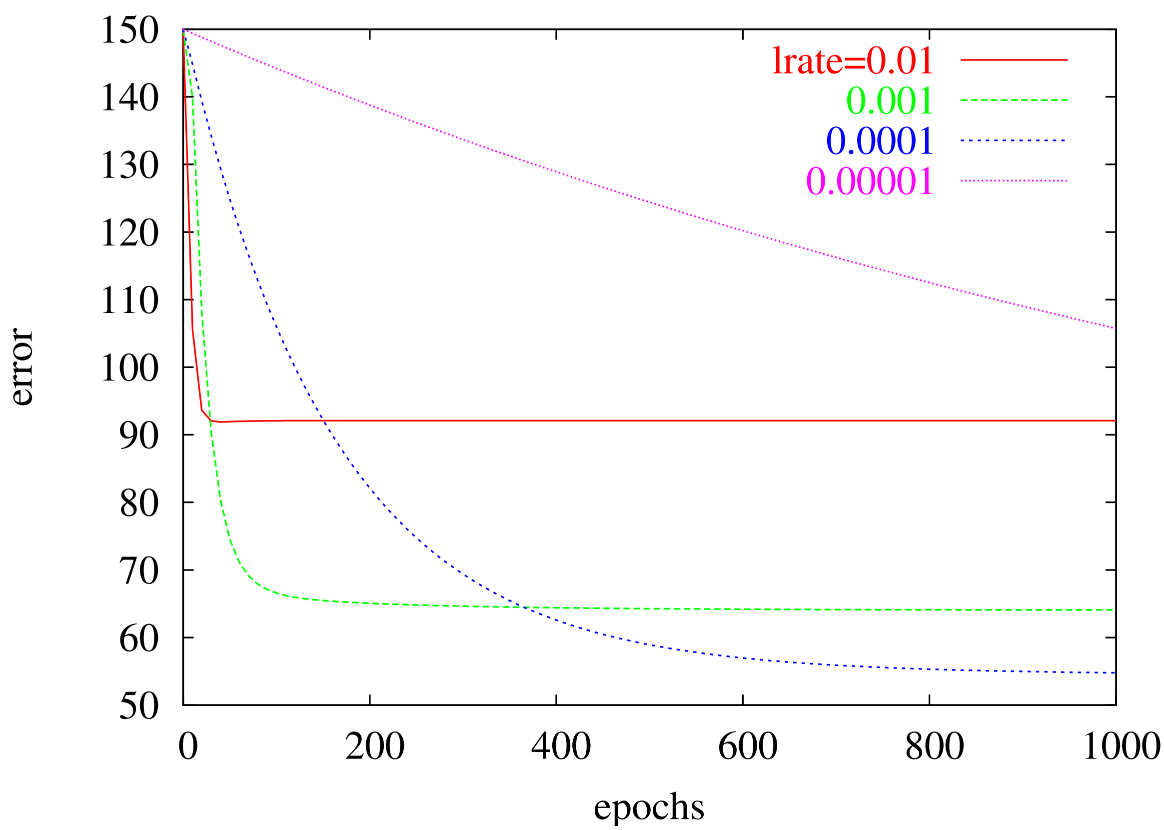
In the first experiment ($1/X^2 \approx 0.0189$) different learning rates were used to learn the third class in iris2d.data (first plot).

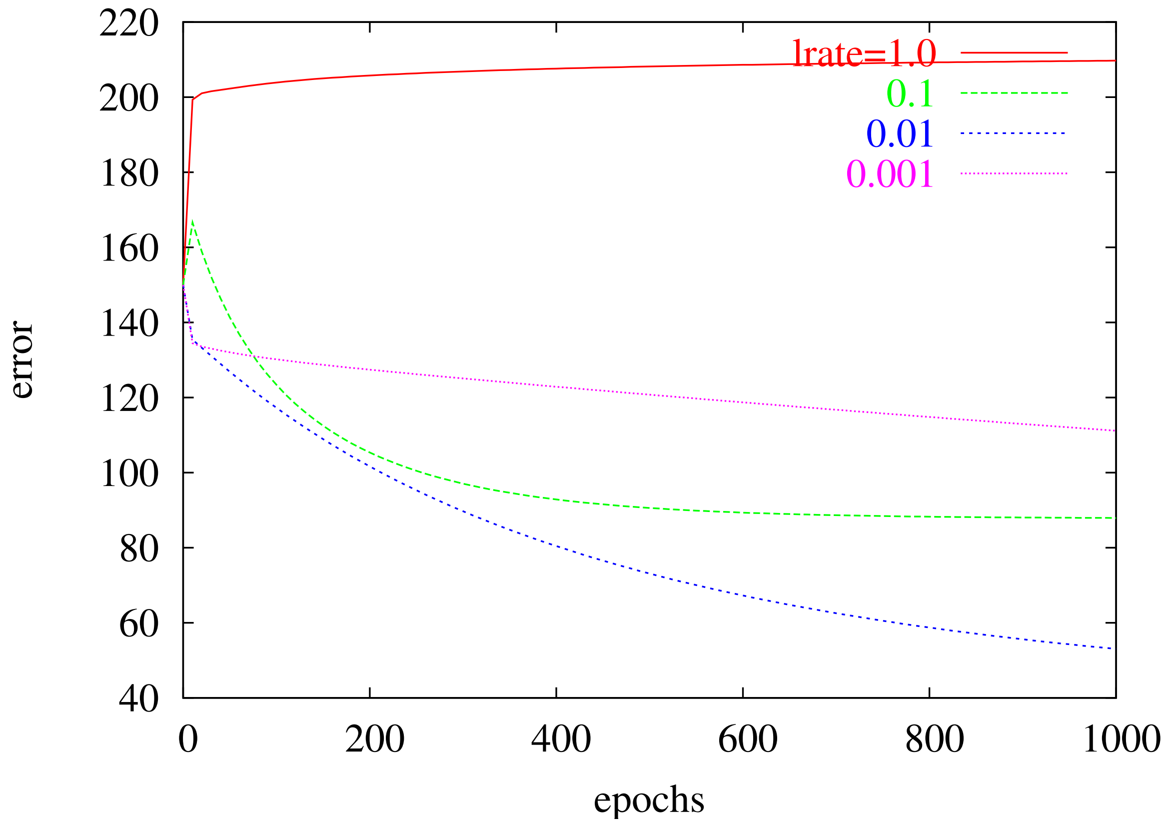
In the second experiment ($1/X^2 \approx 0.485$), I included quadratic terms

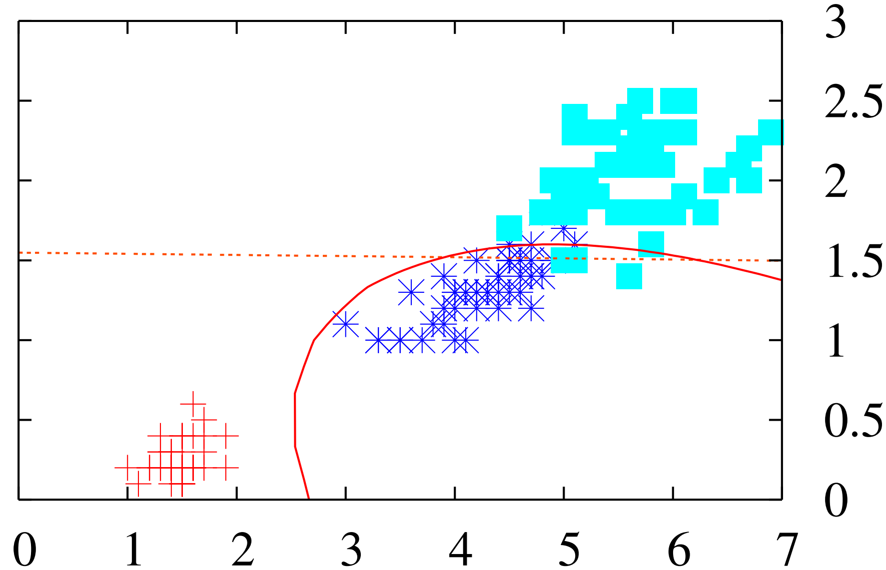
$$\left(x_1/10, x_2/5, (x_1/10)^2, (x_2/5)^2, x_1x_2/50\right)$$

to learn the second class (second plot).

The third plot shows the line and curve for the best learning rates from the two experiements.




lrate=1.0
0.1
0.01
0.001
error
epochs
220
200
180
160
140
120
100
80
60
40
0
200
400
600
800
1000


3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
7