WASP: AIML Module 2

June 2022

1 Logistic Regression

Logistic regression uses an s-shaped function that maps a set up inputs to a value between 0 and 1, defined by:

$$p(y = 1|t) = \frac{e^z}{1 + e^z}$$

where z is a function, of some weights, inputs and biases (w, x, b, respectively), such that

$$z(x) = w \cdot x + b$$

and

$$y(x) = 0 \text{ if } z(x) < .5, \text{ else } 1$$

The loss can be defined as:

$$L(x) = \frac{1}{n} \sum_{i=1}^{n} log\Big(1 + e^{-y_i(x)}\Big)$$

where n is the number of samples. Our objective is to minimize the total loss across all of the samples by changing the parameters w, b. Since this function is 'nice' in the Calculus sense, we can minimize it analytically, using the gradient.

1.1 Gradient

To generalize this further, for any $z(x) = \theta \cdot x$

$$log(p(x)) = log\Big(\frac{1}{1 + e^{-\theta x}}\Big) = -log\Big(1 + e^{-\theta x}\Big)$$

which yields

$$L(x) = \frac{1}{n} \sum_{i=1}^{n} -log\Big(y_i \cdot \theta \cdot x_i - log(1 + e^{-}\theta x)\Big).$$

Therefore

$$\frac{\partial}{\partial \theta_j} y_i \cdot \theta \cdot x_{i,j} = y_i \cdot x_{i,j}$$

and

$$\frac{\partial}{\partial \theta_j} - log(1 + e^{-}\theta x) = x_{j,i} z(x_i)$$

such that

$$\frac{\partial}{\partial \theta_j} = L(\theta) \frac{1}{n} \sum_{i=1}^{m} (z(x^i) - y^i) \cdot x_{i,j}$$

2 Experiments

I looked at both the feature mapping experiments as well as the learning rate experiments.

2.1 Feature Mapping

In this experiment, we used the original feature map to derive two more based on power transformations of the original data using powers 2 and 3 respectively.

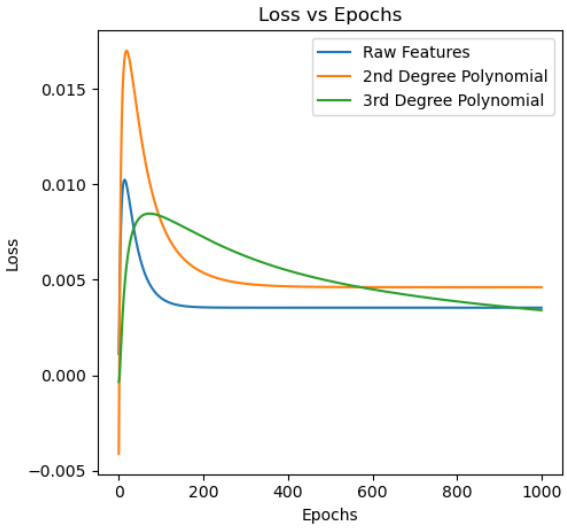

Figure 1: Pictured are three logistic regression models that are trained on various power transformations of the data. As we can see

2.2 Learning Rate

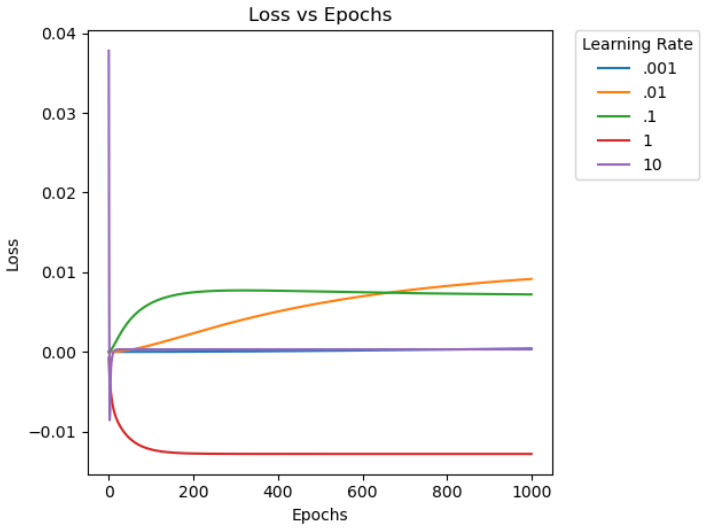

Figure 2: Pictured are three logistic regression models that are trained using various learning rate strategies. As we can see, the raw features and the 3rd degree features converge faster, but they all converge on roughly the same loss, which a slight disadvantage to the 2nd degree transformation.

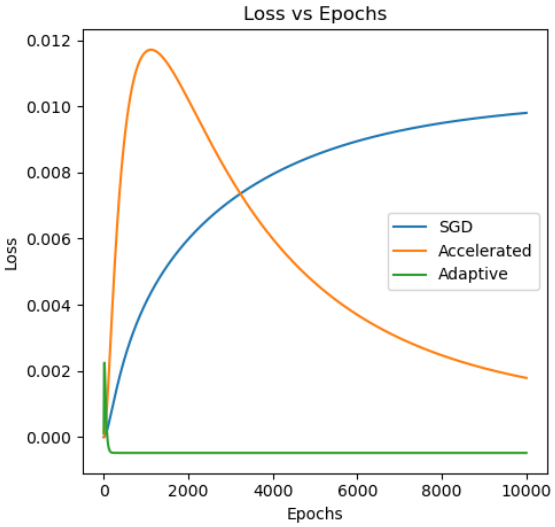

Figure 3: Pictured are three logistic regression models that are trained using various learning rate strategies. As we can see, the adaptive method converges incredibly quickly and nearly to zero while the adaptive method lags behind. The stochastic gradient technique performs the worst, due to the fixed step size preventing a smaller loss.

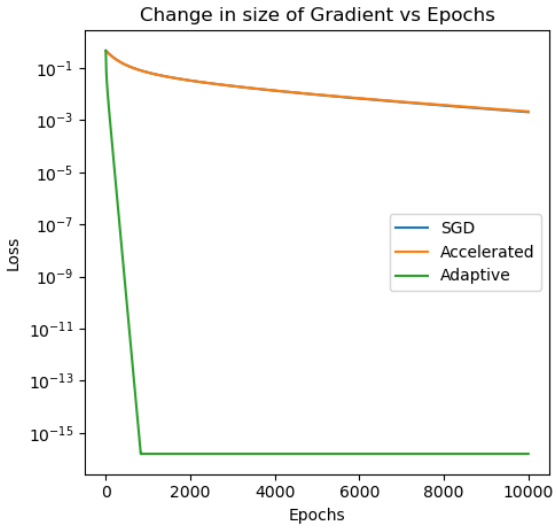

Figure 4: Pictured are three logistic regression models that are trained using various learning rate step sizes. The accelerated method takes marginally less time than the stochastic gradient method with a step size = 10 (as optimized above). The adaptive method, however, is much faster.