ECE 473 Homework 3

Jason Park (park1036)

January 27, 2022

1. Loss function:
$$Loss = (\sigma(\vec{w}\phi(x) - y)^2$$

2. Compute gradient of the loss function with respect to $\vec{w}$
$$p = \sigma(\vec{w}\phi(x))$$
   (a) $\bigtriangledown_{\vec{(w)}} = 2(\sigma(\vec{(w)}\phi(x) - y)\sigma'(\vec{(w)}\phi(x))\phi(x)$
   $= 2\phi(x)\sigma(\vec{(w)}\phi(x) - y)\sigma'(\vec{(w)}\phi(x))$
   (b) $\sigma(z) = (1 + e^{-z})^{-1}$
   (c) $1 + e^{-z} = \sigma^{-1}(z)$
   (d) $e^{-z} = \sigma^{-1}(z) - 1$
   (e) $\sigma'(z) = -1(1 + e^{-z})^{-2}(-e^{-z})$
   $= \sigma^2(z)e^{-z}$
   $= \sigma^2(z)(\sigma^{-1}(z) - 1)$
   (f) $\sigma'(z) = \sigma(z) - \sigma^2(z)$
   (g) $\bigtriangledown_{\vec{(w)}} = 2\phi(x)(p - y)(p - p^2)$
   $= -2\phi(x)p(p - y)(p - 1)$

3. To make $\bigtriangledown_{\vec{(w)}} = -2\phi(x)p(p - y)^2$ arbitrarily small, and given $\sigma(z) = (1 + e^{-z})^{-1}$, the value of $\vec{w}$ would have to approach $-\infty$. The magnitude can never be 0.

4. Find largest magnitude that gradient can take:
   (a) $\bigtriangledown_{\vec{(w)}} = -2\phi(x)p(p - y)^2 = -2\phi(x)p(p^2 - 2p + 1) = -2\phi(x)(p^3 - 2p^2 + p)$
   (b) $(\frac{d}{dp})(-2\phi(x))(p^3 - 2p^2 + p) = -2\phi(x)(3p^2 - 4p + 1) = -2\phi(x)(3p - 1)(p - 1)$
   (c) Largest Magnitude can occur @ $p = \frac{1}{3}$ or $p = 1$
   (d) Plugging in $p = \frac{1}{3}$ and $p = 1$ into $|| - 2\phi(x)p(p - y)^2||$
   $p = \frac{1}{3} : \phi(x)(\frac{8}{27}) = (\frac{8}{27})\phi(x)$
   $p = 1 : \phi(x)(0) = 0$
   (e) Since, the largest magnitude can occur @ $p = \frac{1}{3}$ or $p = 1$, and when $p = 1$ the magnitude is 0, the largest magnitude is $(\frac{8}{27})\phi(x)$, occurring at $p = \frac{1}{3}$.

5. Show that there is an easy transformation to a modified data set D' of (x, y') pairs such that performing least squares regression (using a linear predictor and the squared loss) on D' converges to a vector w that yields zero loss on D'

   (a) $Loss_D = (\sigma(\vec{w}\phi(x)) - y)^2$
   $Loss_{D'} = (\vec{w}\phi(x) - y')^2$

   (b) $\sigma(\vec{w}\phi(x)) - y = 0$

   (c) $\vec{w}'\phi(x) - y' = 0$

   (d) $y = \sigma(\vec{w}\phi(x))$

   (e) $y' = \vec{w}'\phi(x)$

   (f) $\sigma^{-1}(\sigma(\vec{w}\phi(x))) = \sigma^{-1}(y)$

   (g) $\sigma^{-1}(\vec{w'}\phi(x)) = \sigma^{-1}(y')$

   (h) $\vec{w}'\phi(x) = y' = \sigma^{-1}(y)$

6. The first aspect of the code that I noticed first was in the output. My first thought was that between the mean loss of gradient descent and mean loss stochastic gradient descent, there was no observable pattern between the two. The other main aspect of the output that I noticed was that for experiment 2 SGD, loss was much lower than gradient descent. Upon further investigation of the code, I also noticed that SGD does not go through the data after the first pass, especially for a very large set where we cannot afford to go through it many times. All of this is consistent with the properties of SGD vs gradient descent. SGD would be faster, and only needs one iteration, because we only use one sample from the data set to update the parameter(s) for the iteration. This contrasts and is confirmed by the output from the python code, which also shows gradient descent taking a long time, since we have to go through all the samples to make one update iteration.