Homework 5

Michael Parker

Dec 11, 2021

1 Paper Problems [40 points]

1. [5 points] $z = \sigma(y_1^2 + y_2 y_3)$, where $y_1 = 3x$, $y_2 = e^{-x}$, $y_3 = \sin(x)$
   Noting that $y_1(0) = 0, y_2(0) = 1, y_3(0) = 0$,

$$\frac{\partial z}{\partial x} = \frac{\partial}{\partial x}\sigma(y_1^2 + y_2 y_3)$$

$$= \sigma(y_1^2 + y_2 y_3)(1 - \sigma(y_1^2 + y_2 y_3))\frac{\partial}{\partial x}(y_1^2 + y_2 y_3)$$

$$= \sigma(y_1^2 + y_2 y_3)(1 - \sigma(y_1^2 + y_2 y_3))(2y_1\frac{\partial}{\partial x}y_1 + y_2\frac{\partial}{\partial x}y_3 + y_3\frac{\partial}{\partial x}y_2)$$

$$= \sigma(y_1^2 + y_2 y_3)(1 - \sigma(y_1^2 + y_2 y_3))(2y_1(3) + y_2 cos(x) - y_3 e^{-x})$$

$$\left.\frac{\partial z}{\partial x}\right|_0 = \sigma(0)(1 - \sigma(0))(2(0)(3) + 1(1) - 0(1))$$

$$= 0.5 * 0.5 * 1 = 0.25$$

2. [5 points] We'll go layer by layer, forward from 0.

   Layer 0:
   $x_0 = 1$
   $x_1 = 1$
   $x_2 = 1$
   (given)

   Layer 1:
   $z_0^1 = 1$ (constant)
   $z_1^1 = \sigma(w_{01}^1 x_0 + w_{11}^1 x_1 + w_{21}^1 x_2) = \sigma(-1 * 1 - 2 * 1 - 3 * 1) = \sigma(-6) \approx 0.0024726$
   $z_2^1 = \sigma(w_{02}^1 x_0 + w_{12}^1 x_1 + w_{22}^1 x_2) = \sigma(1 * 1 + 2 * 1 + 3 * 1) = \sigma(6) \approx 0.9975274$

   Layer 2:
   $z_0^2 = 1$ (constant)
   $z_1^2 = \sigma(w_{01}^2 z_0^1 + w_{11}^2 z_1^1 + w_{21}^2 z_2^1) = \sigma(-1 * 1 - 2\sigma(-6) - 3\sigma(6)) \approx \sigma(-3.9975274) \approx 0.0180299$

$z_2^2 = \sigma(w_{02}^2 z_0^1 + w_{12}^2 z_1^1 + w_{22}^2 z_2^1) = \sigma(1*1+2\sigma(-6)+3\sigma(6)) \approx \sigma(3.9975274) \approx 0.9819701$

Layer 3:
$y = w_{01}^3 z_0^2 + w_{11}^3 z_1^2 + w_{21}^3 z_2^2 \approx -1 + 2*0.0180299 - 1.5*0.9819701 \approx -2.4368952$

3. [20 points] Layer by layer again, but this time backwards:

Layer 3:
$\frac{\partial L}{\partial y} = (y - y^*) \approx -3.4368952$

$\frac{\partial L}{\partial w_{01}^3} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial w_{01}^3} = \frac{\partial L}{\partial y} z_0^2 \approx -3.4368952 * 1 \approx -3.4368952$
$\frac{\partial L}{\partial w_{11}^3} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial w_{11}^3} = \frac{\partial L}{\partial y} z_1^2 \approx -3.4368952 * 0.0180299 \approx -0.0619670$
$\frac{\partial L}{\partial w_{21}^3} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial w_{21}^3} = \frac{\partial L}{\partial y} z_2^2 \approx -3.4368952 * 0.9819701 \approx -3.3749282$

Layer 2:
$\frac{\partial L}{\partial z_1^2} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial z_1^2} = \frac{\partial L}{\partial y} w_{11}^3 \approx -3.4368952 * 2 \approx -6.8737905$
$\frac{\partial L}{\partial z_2^2} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial z_2^2} = \frac{\partial L}{\partial y} w_{21}^3 \approx -3.4368952 * -1.5 \approx 5.1553428$

Let $s_1^2 = w_{01}^2 z_0^1 + w_{11}^2 z_1^1 + w_{21}^2 z_2^1$, (thus $z_1^2 = \sigma(s_1^2)$), and
$s_2^2 = w_{02}^2 z_0^1 + w_{12}^2 z_1^1 + w_{22}^2 z_2^1$

$\frac{\partial L}{\partial s_1^2} = \frac{\partial L}{\partial z_1^2}\frac{\partial z_1^2}{\partial s_1^2} = \frac{\partial L}{\partial z_1^2}\sigma(s_1^2)(1 - \sigma(s_1^2)) = \frac{\partial L}{\partial z_1^2} z_1^2(1 - z_1^2) \approx -6.8737905 * 0.0180299 * (1 - 0.0180299) \approx -0.1216995$
$\frac{\partial L}{\partial s_2^2} = \frac{\partial L}{\partial z_2^2}\frac{\partial z_2^2}{\partial s_2^2} = \frac{\partial L}{\partial z_2^2}\sigma(s_2^2)(1 - \sigma(s_2^2)) = \frac{\partial L}{\partial z_2^2} z_2^2(1 - z_2^2) \approx 5.1553428 * 0.9819701 * (1 - 0.9819701) \approx 0.0912746$
$\frac{\partial L}{\partial w_{01}^2} = \frac{\partial L}{\partial s_1^2}\frac{\partial s_1^2}{\partial w_{01}^2} = \frac{\partial L}{\partial s_1^2} z_0^1 \approx -0.1216995 * 1 \approx -0.1216995$
$\frac{\partial L}{\partial w_{11}^2} = \frac{\partial L}{\partial s_1^2}\frac{\partial s_1^2}{\partial w_{11}^2} = \frac{\partial L}{\partial s_1^2} z_1^1 \approx -0.1216995 * 0.0024726 \approx -0.0003009$
$\frac{\partial L}{\partial w_{21}^2} = \frac{\partial L}{\partial s_1^2}\frac{\partial s_1^2}{\partial w_{21}^2} = \frac{\partial L}{\partial s_1^2} z_2^1 \approx -0.1216995 * 0.9975274 \approx -0.1213986$
$\frac{\partial L}{\partial w_{02}^2} = \frac{\partial L}{\partial s_2^2}\frac{\partial s_2^2}{\partial w_{02}^2} = \frac{\partial L}{\partial s_2^2} z_0^1 \approx 0.0912746 * 1 \approx 0.0912746$
$\frac{\partial L}{\partial w_{12}^2} = \frac{\partial L}{\partial s_2^2}\frac{\partial s_2^2}{\partial w_{12}^2} = \frac{\partial L}{\partial s_2^2} z_1^1 \approx 0.0912746 * 0.00024726 \approx 0.0022569$
$\frac{\partial L}{\partial w_{22}^2} = \frac{\partial L}{\partial s_2^2}\frac{\partial s_2^2}{\partial w_{22}^2} = \frac{\partial L}{\partial s_2^2} z_2^1 \approx 0.0912746 * 0.9975274 \approx 0.9104892$

Layer 1:
$\frac{\partial L}{\partial z_1^1} = \frac{\partial L}{\partial s_1^2}\frac{\partial s_1^2}{\partial z_1^1} + \frac{\partial L}{\partial s_2^2}\frac{\partial s_2^2}{\partial z_1^1} = \frac{\partial L}{\partial s_1^2} w_{11}^2 + \frac{\partial L}{\partial s_2^2} w_{12}^2 \approx -0.1216995*-2+0.0912746*2 \approx 0.4259482$
$\frac{\partial L}{\partial z_2^1} = \frac{\partial L}{\partial s_1^2}\frac{\partial s_1^2}{\partial z_2^1} + \frac{\partial L}{\partial s_2^2}\frac{\partial s_2^2}{\partial z_2^1} = \frac{\partial L}{\partial s_1^2} w_{21}^2 + \frac{\partial L}{\partial s_2^2} w_{22}^2 \approx -0.1216995*-3+0.0912746*3 \approx 0.6389222$

$s_1^1 = w_{01}^1 x_0 + w_{11}^1 x_1 + w_{21}^1 x_2$, and
$s_2^1 = w_{02}^1 x_0 + w_{12}^1 x_1 + w_{22}^1 x_2$

$\frac{\partial L}{\partial s_1^1} = \frac{\partial L}{\partial z_1^1}\frac{\partial z_1^1}{\partial s_1^1} = \frac{\partial L}{\partial z_1^1}\sigma(s_1^1)(1 - \sigma(s_1^1)) = \frac{\partial L}{\partial z_1^1} z_1^1(1 - z_1^1) \approx 0.4259482 * 0.0024726 * (1 - 0.0024726) \approx 0.0010506$

$\frac{\partial L}{\partial s_2^1} = \frac{\partial L}{\partial z_2^1}\frac{\partial z_2^1}{\partial s_2^1} = \frac{\partial L}{\partial z_2^1}\sigma(s_2^1)(1-\sigma(s_2^1)) = \frac{\partial L}{\partial z_2^1} z_2^1(1-z_2^1) \approx 0.6389222 * 0.9975274 * (1 - 0.0024726) \approx 0.0015759$
$\frac{\partial L}{\partial w_{01}^1} = \frac{\partial L}{\partial s_1^1}\frac{\partial s_1^1}{\partial w_{01}^1} = \frac{\partial L}{\partial s_1^1} x_0 \approx 0.0010506 * 1 \approx 0.0010506$
$\frac{\partial L}{\partial w_{11}^1} = \frac{\partial L}{\partial s_1^1}\frac{\partial s_1^1}{\partial w_{11}^1} = \frac{\partial L}{\partial s_1^1} x_1 \approx 0.0010506 * 1 \approx 0.0010506$
$\frac{\partial L}{\partial w_{21}^1} = \frac{\partial L}{\partial s_1^1}\frac{\partial s_1^1}{\partial w_{21}^1} = \frac{\partial L}{\partial s_1^1} x_2 \approx 0.0010506 * 1 \approx 0.0010506$
$\frac{\partial L}{\partial w_{02}^1} = \frac{\partial L}{\partial s_1^1}\frac{\partial s_1^1}{\partial w_{02}^1} = \frac{\partial L}{\partial s_2^1} x_0 \approx 0.0015759 * 1 \approx 0.0015759$
$\frac{\partial L}{\partial w_{12}^1} = \frac{\partial L}{\partial s_1^1}\frac{\partial s_1^1}{\partial w_{12}^1} = \frac{\partial L}{\partial s_2^1} x_1 \approx 0.0015759 * 1 \approx 0.0015759$
$\frac{\partial L}{\partial w_{22}^1} = \frac{\partial L}{\partial s_1^1}\frac{\partial s_1^1}{\partial w_{22}^1} = \frac{\partial L}{\partial s_2^1} x_2 \approx 0.0015759 * 1 \approx 0.0015759$

4. [10 points]

- [7 points] We want to solve (noting $\sigma = 1$ in our case):

$$\min_w \sum_1^m log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \mathbf{w}^T \mathbf{w}$$

  The gradient of this is:

$$\nabla \sum_1^m log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \mathbf{w}^T \mathbf{w}$$

$$= \sum_1^m \nabla log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \nabla \mathbf{w}^T \mathbf{w}$$

$$\implies \nabla_k = \sum_1^m \frac{-y_{ik} x_{ik} exp(-y_i \mathbf{w}^T \mathbf{x}_i)}{1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)} + 2w_i$$

$$= \sum_1^m -y_{ik} x_{ik} \sigma(y_i \mathbf{w}^T \mathbf{x}_i) + 2w_i$$

- [3 points] We're starting with $w_0 = w_1 = w_2 = w_3 = 0$. Remembering that in SGD we need to multiply the first term by the count of examples (since we're not adding them all up), and that we are using an implied $x_0 = 1$:

  Step 1:
  $\nabla = \{-3*1*1*\sigma(0)+0, -3*1*0.5*\sigma(0)+0, -3*1*-1*\sigma(0)+0, -3*1*0.3*\sigma(0)+0\}$
  $= \{-1.5, -0.75, 1.5, -.45\}$
  $w = \{0.015, 0.0075, -0.015, 0.0045\}$

  Step 2:

$\sigma(y\mathbf{w}^T\mathbf{x}) = \sigma(-(1(0.015) - 1(0.0075) - 2(-0.015) - 2(0.0045))) = \sigma(-0.0285) \approx 0.4928755$
$\nabla = \{-3 * -1 * 1 * 0.4928755 + 0.015, -3 * -1 * -1 * 0.4928755 + 0.0075, -3 * -1 * -2 * 0.4928755 - 0.015, -3 * -1 * -2 * 0.4928755 + 0.0045\}$
$= \{1.4936265, -1.4711265, -2.972253, -2.952753\}$
$w = \{0.015 - 0.005 * 1.4936265, 0.0075 + 0.005 * 1.4711265, -0.015 + 0.005 * 2.972253, 0.0045 + 0.005 * 2.952753\}$
$= \{0.0075318675, 0.0148556325, -0.000138735, 0.019263765\}$

Step 3:
$\sigma(y\mathbf{w}^T\mathbf{x}) = \sigma((1(0.0075318675)+1.5(0.0148556325)+0.2(-0.000138735)-2.5(0.019263765))) = \sigma(-0.01837184325) \approx 0.495407168370$
$\nabla = \{-3 * 1 * 1 * 0.495407168370 + 0.0075318675, -3 * 1 * 1.5 * 0.495407168370 + 0.0148556325, -3*1*0.2*0.495407168370-0.000138735, -3*1*-2.5*0.495407168370+ 0.019263765\}$
$= \{-1.4786896, -2.2144766, -0.2973830, 3.734817527775\}$
$w = \{0.0075318675+0.0025*1.4786896, 0.0148556325+0.0025*2.2144766, -0.000138735+ 0.0025 * 0.2973830, 0.019263765 - 0.0025 * 3.734817527775\}$
$= \{0.0112286, 0.0203918, 0.00060472, 0.0099267\}$

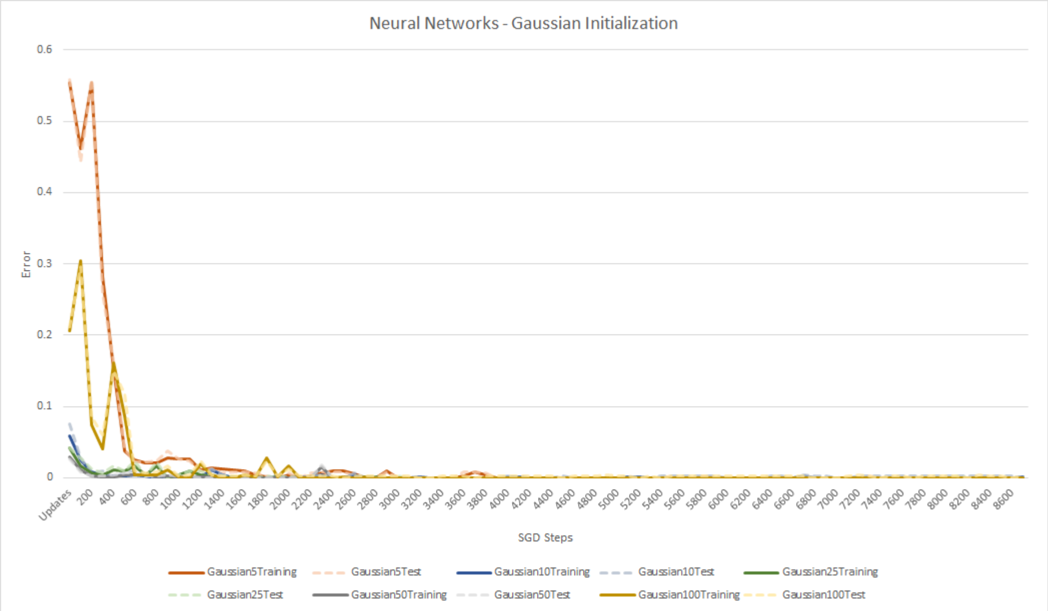
2 Practice [62 points + 60 bonus]

1. [2 Points] Done, pushed to https://github.com/FakerMike/MachineLearningLibrary

2. [58 points]

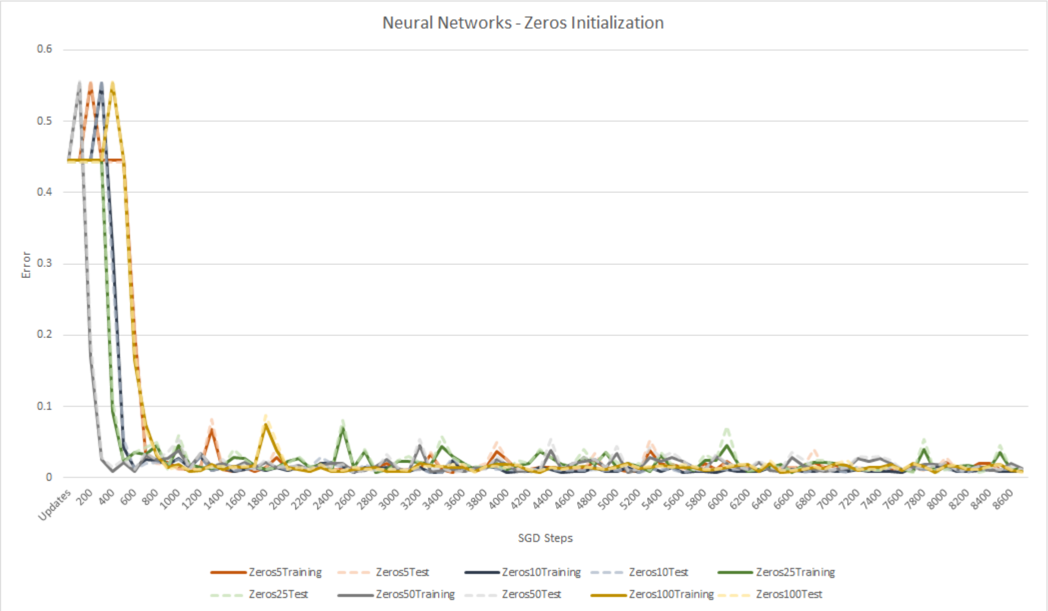
   (a) [25 points] Done, checked against manual calculations and got the same values (they will both print out when running the program).

   (b) [17 points] Ended up seeing pretty good convergence on all of the networks with 10 epochs, $\gamma_o = 0.05$ and $d = 2$:



I noticed that changing γ_0 and d would affect which of the networks would converge first. The bigger (wider) networks tended to wander around a lot more with a higher gamma, taking longer to converge down. Setting it lower ended up making the smaller network also take a while to converge. But overall, all of the networks got to really good test error (either zero or one incorrect test example) within two epochs, and stopped having any deviations from that after five.

(c) [10 points] Using the same settings as above (except for the initialization) gave this:



Regardless of the width of the network, they all took pretty much the same path to convergence (albeit at slightly different rates). When they did converge, they had much higher error (usually about 0.012) than the ones initialized randomly. This makes sense, because by initializing them all to zero, all their derivatives will be identical for each node in a layer. This means we have essentially reduced our structure's dimension to our depth (and faced the corresponding drop in expressiveness). So it's not surprising we ended up with more or less identical performance to our linear classifiers - the neural network isn't quite linear (due to the sigmoids), but it is only a 3 dimensional object and thus can't find a perfect fit.

(d) [6 points] The neural network has really good expressiveness even in its simpler form. With only 2 hidden layers and width 5, it was able to get to 0 training and 0 test error. It's also very very quick even though it is a nonlinear classifier - running thousands of updates on the smaller networks was done in an instant.

(e) [**Bonus**] [30 points]

3. [**Bonus**] [30 points]

   (a) [10 points]

   (b) [5 points]

   (c) [3 points]

4. [2 Points] Uploaded. I'm pretty proud of the library, especially of the AdaBoost and the Neural Networks implementations - these two are powerful classifiers and yet fit neatly enough into my code that I could easily rebuild them later.