hw_5

Matthew Jensen

2024-12-02

Part 1

Problem 1.

z = $\sigma(y_1^2 + y_2 y_3)$

$y_1 = 3x$

$y_2 = e^{-x}$

$y_3 = sin(x)$

$\sigma(v) = \frac{1}{1+e^{-v}}$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial v} \cdot \frac{\partial v}{\partial x}$$

$$\frac{\partial z}{\partial v} = \sigma(v)(1 - \sigma(v))$$

$$\text{For } v = y_1^2 + y_2 y_3, \frac{\partial v}{\partial x} = \frac{\partial (y_1^2)}{\partial x} + \frac{\partial (y_2 \cdot y_3)}{\partial x}$$

$$\frac{\partial (y_1^2)}{\partial x} = \frac{\partial}{\partial x} 9x^2 = 18x$$

$$\frac{\partial (y_2 \cdot y_3)}{\partial x} = e^{-x} \cdot cos(x) - e^{-x}(sin(x))$$

$$\text{Thus, } \frac{\partial v}{\partial x} = 18x + (e^{-x} \cdot cos(x) - e^{-x}(sin(x)))$$

$$\text{Thus, } \frac{\partial z}{\partial x} = (\sigma(v)(1 - \sigma(v))) \cdot (18x + (e^{-x} \cdot cos(x) - e^{-x}(sin(x)))) \text{ ,where } v = y_1^2 + y_2 y_3$$

Evaluate at x = 0

$$v = 3(0) + (e^0 \cdot sin(0)) = 0 + 1(0) = 0, \sigma(0) = \frac{1}{1 + e^0} = \frac{1}{2}$$

$$\frac{\partial v}{\partial x} = 18(0) + (e^o \cdot cos(0) - e^0(sin(0))) = 0 + ((1 \cdot 1) + 0) = 1$$

$$\text{Thus, } \frac{\partial z}{\partial x} = (\frac{1}{2}(1 - \frac{1}{2})) \cdot 1 = \mathbf{\frac{1}{4}}$$

Problem 2.

Layer 1 Computations

$x_0 = x_1 = x_2 = 1$

Using these 3 values, we will calculate z_1^1 and z_2^1

$$z_1^1 = \sigma(w_{01}^1 x_0 + w_{11}^1 x_1 + w_{21}^1 x_2)$$

$$= \sigma(-1(1) + -2(1) + -3(1)) = \sigma(-6) = .00247$$

$$z_2^1 = \sigma(w_{02}^1 x_0 + w_{12}^1 x_1 + w_{22}^1 x_2)$$

$$= \sigma(1(1) + 2(1) + 3(1)) = \sigma(6) = .9975$$

Layer 2 Computations

z_0^1 = 1

z_1^1 = .00247

z_2^1 = .9975

$$z_1^2 = \sigma(w_{01}^2 z_0^1 + w_{11}^2 z_1^1 + w_{21}^2 z_2^1)$$

$$= \sigma(-1(1) + -2(.00247) + -3(.9975)) = \sigma(-3.99744) = .018$$

$$z_2^2 = \sigma(w_{02}^2 z_0^1 + w_{12}^2 z_1^1 + w_{22}^2 z_2^1)$$

$$= \sigma(1(1) + 2(.00247) + 3(.9975)) = \sigma(3.99744) = .982$$

Layer 3 Computations (0utput Layer)

$$y = \sigma(w_{01}^3 z_0^2 + w_{11}^3 z_1^2 + w_{21}^3 z_2^2)$$

$$= \sigma(-1(1) + 2(.018) + -1.5(.982)) = \sigma(-2.437) = \mathbf{.08}$$

Problem 3: Backpropagation

Given Information

We use a three-layer feed-forward neural network with weights provided in Table 1. The input vector is x = [1, 1, 1], and the label is y* = 1. The activation function is the sigmoid function.

Forward Pass

Layer 1

Using weights:

$$W^1 = \begin{bmatrix} -1 & -2 & -3 \\ 1 & 2 & 3 \end{bmatrix}$$

Compute:

$$z_1^1 = \sigma(W_{01}^1 x_0 + W_{11}^1 x_1 + W_{21}^1 x_2) = \sigma(-6) = 0.00247$$

$$z_2^1 = \sigma(W_{02}^1 x_0 + W_{12}^1 x_1 + W_{22}^1 x_2) = \sigma(6) = 0.9975$$

Layer 2

Using weights:

$$W^2 = \begin{bmatrix} -1 & -2 & -3 \\ 1 & 2 & 3 \end{bmatrix}$$

Compute:

$$z_1^2 = \sigma(W_{01}^2 z_0^1 + W_{11}^2 z_1^1 + W_{21}^2 z_2^1) = \sigma(-3.99744) = 0.018$$

$$z_2^2 = \sigma(W_{02}^2 z_0^1 + W_{12}^2 z_1^1 + W_{22}^2 z_2^1) = \sigma(3.99744) = 0.982$$

Output Layer (Layer 3)

Using weights:

$$W^3 = \begin{bmatrix} -1 \\ 2 \\ -1.5 \end{bmatrix}$$

Compute:

$$y = \sigma(W_{01}^3 z_0^2 + W_{11}^3 z_1^2 + W_{21}^3 z_2^2) = \sigma(-2.437) = 0.08$$

Loss Function

The loss function is:

$$L(y, y^*) = \frac{1}{2}(y - y^*)^2$$

Substitute values:

$$L(0.08, 1) = \frac{1}{2}(0.08 - 1)^2 = 0.4232$$

Backpropagation

Output Layer (Layer 3)

$$\frac{\partial L}{\partial y} = y - y^* = 0.08 - 1 = -0.92$$

$$\frac{\partial y}{\partial z_3} = \sigma'(z_3) = \sigma(z_3)(1 - \sigma(z_3)) = 0.08 \times (1 - 0.08) = 0.0736$$

$$\frac{\partial L}{\partial z_3} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial z_3} = -0.92 \cdot 0.0736 = -0.0677$$

Gradients for weights in Layer 3:

$$\frac{\partial L}{\partial W_{01}^3} = \frac{\partial L}{\partial z_3} \cdot z_0^2 = -0.0677 \cdot 1 = -0.0677$$

$$\frac{\partial L}{\partial W_{11}^3} = \frac{\partial L}{\partial z_3} \cdot z_1^2 = -0.0677 \cdot 0.018 = -0.0012$$

$$\frac{\partial L}{\partial W_{21}^3} = \frac{\partial L}{\partial z_3} \cdot z_2^2 = -0.0677 \cdot 0.982 = -0.0665$$

Layer 2

Backpropagate to Layer 2:

$$\frac{\partial z_3}{\partial z_2^j} = W_{j1}^3 \cdot \sigma'(z_2^j)$$

Compute for $j = 1, 2$:

$$\frac{\partial L}{\partial z_2^1} = \frac{\partial L}{\partial z_3} \cdot W_{11}^3 \cdot \sigma'(z_2^1) = -0.0677 \cdot 2 \cdot 0.018 \times (1 - 0.018) = -0.0024$$

$$\frac{\partial L}{\partial z_2^2} = \frac{\partial L}{\partial z_3} \cdot W_{21}^3 \cdot \sigma'(z_2^2) = -0.0677 \cdot -1.5 \cdot 0.982 \times (1 - 0.982) = 0.0018$$

Gradients for weights in Layer 2:

$$\frac{\partial L}{\partial W_{01}^2} = -0.0024 \cdot 1 = -0.0024$$

$$\frac{\partial L}{\partial W_{11}^2} = -0.0024 \cdot 0.00247 = -0.000005928$$

$$\frac{\partial L}{\partial W_{21}^2} = -0.0024 \cdot 0.9975 = -0.002394$$

$$\frac{\partial L}{\partial W_{02}^2} = 0.0018 \cdot 1 = 0.0018$$

$$\frac{\partial L}{\partial W_{12}^2} = 0.0018 \cdot 0.00247 = 0.000004446$$

$$\frac{\partial L}{\partial W_{22}^2} = 0.0018 \cdot 0.9975 = 0.0017955$$

Layer 1

Repeat backpropagation process to compute:

$$\frac{\partial L}{\partial W^1_{01}} = -0.000005928 \cdot 1 = -0.000005928$$

$$\frac{\partial L}{\partial W^1_{11}} = -0.000005928 \cdot 1 = -0.000005928$$

$$\frac{\partial L}{\partial W^1_{21}} = -0.000005928 \cdot 1 = -0.000005928$$

$$\frac{\partial L}{\partial W^1_{02}} = 0.000004446 \cdot 1 = 0.000004446$$

$$\frac{\partial L}{\partial W^1_{12}} = 0.000004446 \cdot 1 = 0.000004446$$

$$\frac{\partial L}{\partial W^1_{22}} = 0.000004446 \cdot 1 = 0.000004446$$

Final Results

All gradients are as follows:

Layer 3:

$$\frac{\partial L}{\partial W^3_{01}} = -0.0677, \frac{\partial L}{\partial W^3_{11}} = -0.0012, \frac{\partial L}{\partial W^3_{21}} = -0.0665$$

Layer 2:

$$\frac{\partial L}{\partial W^2_{01}} = -0.0024, \frac{\partial L}{\partial W^2_{11}} = -0.000005928, \frac{\partial L}{\partial W^2_{21}} = -0.002394$$

$$\frac{\partial L}{\partial W^2_{02}} = 0.0018, \frac{\partial L}{\partial W^2_{12}} = 0.000004446, \frac{\partial L}{\partial W^2_{22}} = 0.0017955$$

Layer 1:

$$\frac{\partial L}{\partial W^1_{01}} = -0.000005928, \frac{\partial L}{\partial W^1_{11}} = -0.000005928, \frac{\partial L}{\partial W^1_{21}} = -0.000005928$$

$$\frac{\partial L}{\partial W^1_{02}} = 0.000004446, \frac{\partial L}{\partial W^1_{12}} = 0.000004446, \frac{\partial L}{\partial W^1_{22}} = 0.000004446$$

4. Logistic Regression with MAP Estimation

Problem Setup

We are tasked with learning a logistic regression model using the training dataset provided in Table 2:

x1	x2	x3	y
0.5	-1	0.3	1
-1	-2	-2	-1
1.5	0.2	-2.5	1

Prior Distribution

Each parameter w_i comes from a standard Gaussian prior distribution:

$$p(w_i) = \mathcal{N}(w_i \mid 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} w_i^2\right).$$

MAP Objective Function

The MAP objective function is the log joint probability:

$$\log P(w \mid D) = \sum_{i=1}^{n} \Big[y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i)) \Big] - \frac{1}{2} \sum_{j=0}^{3} w_j^2,$$

where $z_i = w_0 + w_1 x_{1i} + w_2 x_{2i} + w_3 x_{3i}$ and $\sigma(z) = \frac{1}{1+e^{-z}}$.

Gradient Derivation

The gradient of the objective function with respect to each parameter w_j is:

$$\frac{\partial \log P(w \mid D)}{\partial w_j} = \sum_{i=1}^{n} (y_i - \sigma(z_i)) x_{ij} - w_j,$$

where x_{ij} includes the bias term ($x_{i0} = 1$).

Stochastic Gradient Descent (SGD)

We calculate the stochastic gradient of the objective for the first three steps of SGD.

Step 1: First Example

Data Point:

$x = [0.5, -1, 0.3], \ y = 1$

Computation:

$$z = w_0 + w_1(0.5) + w_2(-1) + w_3(0.3) = 0 \quad \text{(initial weights are all zero).}$$

$$\sigma(z) = 0.5, \quad y - \sigma(z) = 1 - 0.5 = 0.5.$$

Gradients:

$$\frac{\partial \log P}{\partial w_0} = 0.5(1) - 0 = 0.5, \quad \frac{\partial \log P}{\partial w_1} = 0.5(0.5) - 0 = 0.25, \frac{\partial \log P}{\partial w_2} = 0.5(-1) - 0 = -0.5, \frac{\partial \log P}{\partial w_3} = 0.5(0.3) - 0 = 0.15.$$

Step 2: Second Example

Data Point:

$x = [-1, -2, -2], \ y = -1$

Updated Weights:

$$w_0^{(1)} = 0 + 0.01(0.5) = 0.005, \quad w_1^{(1)} = 0.0025, \quad w_2^{(1)} = -0.005, \quad w_3^{(1)} = 0.0015.$$

Computation:

$$z = 0.005 + (-0.0025)(-1) + (-0.005)(-2) + 0.0015(-2) = 0.014.$$

$$\sigma(z) = 0.5035, \quad y - \sigma(z) = -1 - 0.5035 = -1.5035.$$

Gradients:

$$\frac{\partial \log P}{\partial w_0} = -1.5035(1) - 0.005 = -1.5085, \quad \frac{\partial \log P}{\partial w_1} = -1.5035(-1) - 0.0025 = 1.501, \frac{\partial \log P}{\partial w_2} = -1.5035(-2) - (-0.005) = 3.012, \frac{\partial \log P}{\partial w_3} = -1.5035(-2) - 0.0015$$

Step 3: Third Example

Data Point:

$x = [1.5, 0.2, -2.5], \ y = 1$

Updated Weights:

$$w_0^{(2)} = 0.005 + 0.005(-1.5085) = -0.0025, \quad w_1^{(2)} = 0.0025 + 0.005(1.501) = 0.01,$$

$$w_2^{(2)} = -0.005 + 0.005(3.012) = 0.01, \quad w_3^{(2)} = 0.0015 + 0.005(3.0065) = 0.0165.$$

Computation:

$$z = -0.0025 + 0.01(1.5) + 0.01(0.2) + 0.0165(-2.5) = -0.03675.$$

$$\sigma(z) = 0.4908, \quad y - \sigma(z) = 1 - 0.4908 = 0.5092.$$

Gradients:

$$\frac{\partial \log P}{\partial w_0} = 0.5092(1) - (-0.0025) = 0.5117, \quad \frac{\partial \log P}{\partial w_1} = 0.5092(1.5) - 0.01 = 0.7538, \frac{\partial \log P}{\partial w_2} = 0.5092(0.2) - 0.01 = 0.0918, \frac{\partial \log P}{\partial w_3} = 0.5092(-2.5) - 0.0165 = -1.2898.$$

Part 2

Part 2.a

See code for output

Part 2.c

2c answer: Zero initialization results in significantly higher training and test errors compared to random initialization. This is probably due to the fact that each neuron starts out the same.

Part 2.d

2d: The Neural networks outperformed SVM and logistic regression, due to them being able to handle non-linear relationships better. However, they require careful tuning of hyperparameters and architecture to achieve optimal results, while SVM and logistic regression are simpler to train.

Part 2.e

2e answer: The ReLU activation with He initialization consistently achieved lower errors(though both performed really well) and faster convergence compared to Tanh with Xavier initialization. It had better performance than the other models above, probably due to the built in Adam optimizer.

Part 3.a and 3.b

See code for output

Part 3.c

3d: The MAP estimation performed better than the ML estimation in both training and testing. This is due to the MAP inherintley having a regularization penalty due to the presence of a prior.The hyperparameter v in MAP is analogous toC in SVM, as both regulate model complexity—smaller values enforce stronger regularization, while larger values allow more flexibility. MAP offers a principled Bayesian approach to regularization, whereas C in SVM is an empirical tuning parameter.