Machine Learning HW1

韋詠欣

Due: September 10, 2025

Problem 1

We consider stochastic gradient descent (SGD) to learn the model

$$h(x_1, x_2) = \sigma(b + w_1 x_1 + w_2 x_2),$$

where σ is the sigmoid function

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

We are given one data point $(x_1, x_2, y) = (1, 2, 3)$ and initial parameter

$$\theta^0 = (b, w_1, w_2) = (4, 5, 6).$$

Loss function:

$$L(\theta) = \frac{1}{2}(h - y)^2.$$

Let $z = b + w_1 x_1 + w_2 x_2$ and $h = \sigma(z)$. The derivative of σ is

$$\begin{aligned}
\sigma'(z) &= \frac{d}{dz}\left(\frac{1}{1 + e^{-z}}\right) \\
&= \frac{-1\,(-e^{-z})}{(1 + e^{-z})^2} \\
&= e^{-z}\left(\frac{1}{1 + e^{-z}}\right)^2 \\
&= \big(1 + e^{-z} - 1\big)\left(\frac{1}{1 + e^{-z}}\right)^2 \\
&= \left(\frac{1}{\sigma(z)} - 1\right)(\sigma(z))^2 \\
&= \sigma(z)\big(1 - \sigma(z)\big).
\end{aligned}$$

Then the gradient of L with respect to θ is

$$\nabla_\theta L = (h - y)\,\sigma'(z)\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}.$$

The SGD update rule (with learning rate η) is

$$\theta^1 = \theta^0 - \eta \nabla_\theta L.$$

Explicitly,

$$\begin{aligned} b^{(1)} &= b^{(0)} - \eta(h-y)\sigma'(z), \\ w_1^{(1)} &= w_1^{(0)} - \eta(h-y)\sigma'(z)\, x_1, \\ w_2^{(1)} &= w_2^{(0)} - \eta(h-y)\sigma'(z)\, x_2. \end{aligned}$$

Substitution of numbers

- $z = 4 + 5 \cdot 1 + 6 \cdot 2 = 21,$
- $h = \sigma(21) = \dfrac{1}{1+e^{-21}},$
- $\sigma'(21) = \sigma(21)\big(1 - \sigma(21)\big).$

Thus,

$$\boxed{\theta^1 = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - \eta\,\big(\sigma(21) - 3\big)\,\sigma(21)\big(1 - \sigma(21)\big) \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}}$$

Problem 2

(a) Derivatives of the sigmoid

$\sigma(x) = \dfrac{1}{1+e^{-x}}$. Then

$$\begin{aligned} \sigma'(x) &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) \\ &= \frac{-1\,(-e^{-x})}{(1+e^{-x})^2} \\ &= e^{-x}\left(\frac{1}{1+e^{-x}}\right)^2 \\ &= \big(1 + e^{-x} - 1\big)\left(\frac{1}{1+e^{-x}}\right)^2 \\ &= \left(\frac{1}{\sigma(x)} - 1\right)(\sigma(x))^2 \\ &= \sigma(x)\big(1 - \sigma(x)\big). \end{aligned}$$

$$
\begin{aligned}
\sigma''(x) &= \frac{d}{dx}\left(\sigma'(x)\right) \\
&= \frac{d}{dx}\left(\sigma(x)(1-\sigma(x))\right) \\
&= \sigma'(x)(1-\sigma(x)) + \sigma(x)(1-\sigma(x))' \\
&= \left(\sigma(x)(1-\sigma(x))\right)(1-\sigma(x)) + \sigma(x)(-\sigma'(x)) \\
&= \left(\sigma(x)(1-\sigma(x))\right)(1-\sigma(x)) + \sigma(x)(-\sigma(x)(1-\sigma(x))) \\
&= \sigma(x)(1-\sigma(x))(1-\sigma(x)-\sigma(x)) \\
&= \sigma(x)(1-\sigma(x))(1-2\sigma(x))
\end{aligned}
$$

$$
\begin{aligned}
\sigma'''(x) &= \frac{d}{dx}\left(\sigma''(x)\right) \\
&= \frac{d}{dx}\left(\sigma(x)(1-\sigma(x))(1-2\sigma(x))\right) \\
&= \frac{d}{dx}\big(2(\sigma(x))^3 - 3(\sigma(x))^2 + \sigma(x)\big) \\
&= 6(\sigma(x))^2(\sigma'(x)) - 6(\sigma(x))(\sigma'(x)) + \sigma'(x) \\
&= (\sigma'(x))(6(\sigma(x))^2 - 6\sigma(x) + 1) \\
&= (\sigma(x)(1-\sigma(x)))(6(\sigma(x))^2 - 6\sigma(x) + 1)
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
\sigma' &= \sigma(1-\sigma) \\
\sigma'' &= \sigma(1-\sigma)(1-2\sigma) \\
\sigma''' &= \sigma(1-\sigma)(6\sigma^2 - 6\sigma + 1)
\end{aligned}
$$

(b) Relation between sigmoid function and hyperbolic tangent

Recall

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}$$

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

Therefore,

$$\begin{aligned} \sigma(2x) &= \frac{e^{2x}}{e^{2x} + 1} \\ 1 - \sigma(2x) &= \frac{1}{e^{2x} + 1} \end{aligned}$$

$$\begin{aligned} \Rightarrow \tanh(x) &= \sigma(2x) - (1 - \sigma(2x)) \\ &= 2\sigma(2x) - 1 \end{aligned}$$

Problem 3

- Why does the sigmoid saturate for very large positive or negative inputs, and how does this cause vanishing gradients?
- Between squared error and cross-entropy loss, which is more appropriate when using sigmoid for binary classification, and why?