111652026 劉馥瑞

(1)

For layer $l \geq 2$, $z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}$, $a^{[l]} = \sigma(z^{[l]})$ and $a^{[1]} = x$.

We want the row vector $(1 \times n_1)$,

$$g^{[1]} = \frac{\partial a^{[L]}}{\partial a^{[1]}} = \nabla_x a^{[L]}(x).$$

Define for each layer the row vector,

$$g^{[l]} = \frac{\partial a^{[L]}}{\partial a^{[l]}} \quad (\text{shape} = 1 \times n_l)$$

Because $a^{[L]}$ is scalar, $g^{[L]} = 1$ (a 1×1 identity).

Using the chain rule,

$$\frac{\partial a^{[l]}}{\partial a^{[l-1]}} = \text{diag}(\sigma'(z^{[l]}))\, W^{[l]} \quad (n_l \times n_{l-1})$$

Hence the backward recursion is

$g^{[l-1]} = g^{[l]} \text{diag}(\sigma'(z^{[l]}))\, W^{[l]}$ for $l = L, L-1, \ldots, 2$,

and the desired gradient is $g^{[1]}$ (a $1 \times n_1$ row).

Because $a^{[1]} = x$, there is no activation derivative at layer 1 — the recursion stops at $l = 2$.

(2)

Why can MLE be biased in small samples even though it is asymptotically unbiased?