Problem Set 3 – Loss Functions and Fitting Models

DS542 – DL4DS

Spring, 2025

Youran Geng

Note: Refer to the equations in the *Understanding Deep Learning* textbook to solve the following problems.

Problem 5.9

Consider a multivariate regression problem in which we predict the height of an individual in meters and their weight in kilos from some data x. Here, the units take quite different values. What problems do you see this causing? Propose two solutions to these problems.

Solution. First problem is that they have disparate influence to the outputs (predictions). Heights typically range from 1-2 meters, while weights typically range from 40 to 150 kilograms. Such huge differences in value and range results in quite disparate weights to their corresponding variables.

This problem may also cause the model hard to converge and poor in performance, since the model may cater to weights when training.

One fix for this is normalization. Normalize each variable (heights and weights) by $z_i = \dfrac{(x_i - \mu_i)}{\sigma_i}$ so that they equally contribute to the loss.

Another fix is to calculate the losses for heights and weights respectively, and calculate a weighted total loss:

$$L = w_1 L_{\text{height}} + w_2 L_{\text{weight}}.$$

□

Problem 6.6

Which of the functions in Figure 6.11 from the book is convex? Justify your answer. Characterize each of the points 1–7 as (i) a local minimum, (ii) the global minimum, or (iii) neither.

Solution. Only function (b) is convex. On function (a), a line segment connecting point 1 and 2 intersects with the curve; on function (c), a line segment connecting point 6 and 7 intersects with the curve.

On function (a), point 1 and 3 are local minima, and point 2 is the global minimum (therefore also a local minimum).

On function (b), point 5 is the global minimum (thus a local minimum) and point 4 is neither.

On function (c), point 6 is the global minimum (thus a local minimum) and point 7 is neither. □

Problem 6.10

Show that the momentum term m_t (equation (6.11)) is an infinite weighted sum of the gradients at the previous iterations and derive an expression for the coefficients (weights) of that sum.

Proof. Let the gradient $\sum_{i \in \mathcal{B}_t} \frac{\partial}{\partial \phi} \ell_i(\phi_t) = g_t$ and define $\mathbf{m}_0 = g_0$. Thus for any $t \geq 1$,

$$\begin{aligned}
\mathbf{m}_{t+1} &= \beta \mathbf{m}_t + (1-\beta) g_t \\
&= \beta(\beta \mathbf{m}_{t-1} + (1-\beta) g_{t-1}) + (1-\beta) g_t \\
&= \beta^2 \mathbf{m}_{t-1} + \beta(1-\beta) g_{t-1} + (1-\beta) g_t \\
&\vdots \\
&= \beta^{t+1} \mathbf{m}_0 + \beta^t (1-\beta) g_1 + \cdots + (1-\beta) g_t \\
&= \beta^{t+1} g_0 + \beta^t (1-\beta) g_1 + \cdots + (1-\beta) g_t.
\end{aligned}$$

Note that the weights sum to 1, since

$$1 + \beta + \beta^2 + \cdots + \beta^t = \frac{1 - \beta^{t+1}}{1 - \beta}$$

given $0 < \beta < 1$. □