Problem Set 3 Loss Functions and Fitting Models

DS542 DL4DS

Spring, 2025

Sicheng Yi (Tiger Yi)

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.
Answer:
Problems: Difference in variance between Height (1.5 - 2.0 m) and Weight (50 - 100 kg). The variation in weight is a lot larger than height, this will lead to class unbalance in training. Also the weight will have a lot more influence in the loss function than height.
Solutions: standardize the data as in $y = \frac{x-\mu}{\sigma}$ where μ is the mean and σ is the standard deviation. Also for the loss function we can scale the weight different than the height: $L = \lambda_1(h - \hat{h})^2 + \lambda_1(w - \hat{w})^2$

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.

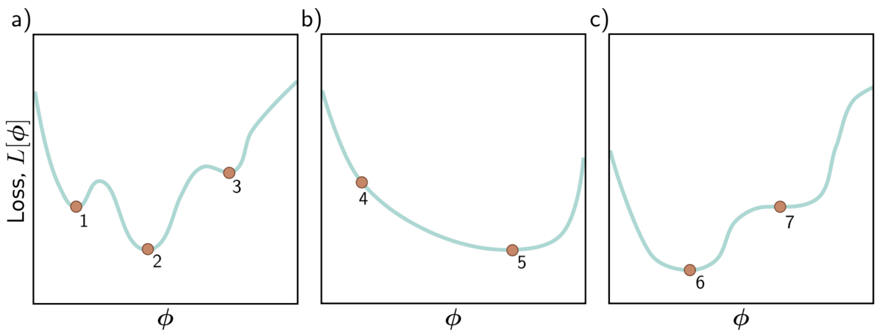

Figure 6.11 Three 1D loss functions for problem 6.6.

Figure 1: problem 6.6

Answer:
Part 1: Only 2nd graph (b) is convex. (a) and (c) are not convex. A convex function satisfies the property that a line segment connecting any two points on the function lies above or on the curve, and only (b) satisfy.
Part 2:
point 1 is local minimum
point 2 is global minimum
point 3 is local minimum
point 4 is neither
point 5 is global minimum
point 6 is global minimum
point 7 is neither (saddle point)

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.

$$\begin{aligned} \mathbf{m}_{t+1} &\leftarrow \beta \cdot \mathbf{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}} \\ \boldsymbol{\phi}_{t+1} &\leftarrow \boldsymbol{\phi}_t - \alpha \cdot \mathbf{m}_{t+1}, \end{aligned} \tag{6.11}$$

where $\mathbf{m}_t$ is the momentum (which drives the update at iteration t), $\beta \in [0, 1)$ controls the degree to which the gradient is smoothed over time, and α is the learning rate.

Figure 2: problem 6.10

Answer:

$$\mathbf{m}_{t+1} = \beta \mathbf{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

Expanding $\mathbf{m}_t$ recursively:

$$\mathbf{m}_t = \beta \mathbf{m}_{t-1} + (1-\beta) \sum_{i \in \mathcal{B}_{t-1}} \frac{\partial \ell_i[\boldsymbol{\phi}_{t-1}]}{\partial \boldsymbol{\phi}}$$

Substituting $\mathbf{m}_{t-1}$:

$$\mathbf{m}_t = \beta \left(\beta \mathbf{m}_{t-2} + (1-\beta) \sum_{i \in \mathcal{B}_{t-2}} \frac{\partial \ell_i[\boldsymbol{\phi}_{t-2}]}{\partial \boldsymbol{\phi}} \right) + (1-\beta) \sum_{i \in \mathcal{B}_{t-1}} \frac{\partial \ell_i[\boldsymbol{\phi}_{t-1}]}{\partial \boldsymbol{\phi}}$$

$$= \beta^2 \mathbf{m}_{t-2} + (1-\beta) \sum_{i \in \mathcal{B}_{t-2}} \beta \frac{\partial \ell_i[\boldsymbol{\phi}_{t-2}]}{\partial \boldsymbol{\phi}} + (1-\beta) \sum_{i \in \mathcal{B}_{t-1}} \frac{\partial \ell_i[\boldsymbol{\phi}_{t-1}]}{\partial \boldsymbol{\phi}}$$

Continuing this expansion back to $\mathbf{m}_0 = 0$ (assuming zero initialization):

$$\mathbf{m}_t = (1-\beta) \sum_{k=0}^{t} \beta^{t-k} \sum_{i \in \mathcal{B}_k} \frac{\partial \ell_i[\boldsymbol{\phi}_k]}{\partial \boldsymbol{\phi}}$$

This equation shows that $\mathbf{m}_t$ is a weighted sum of all past gradients, where the weight assigned to the gradient at iteration k is $w_k = (1-\beta)\beta^{t-k}$.

These weights form a decaying geometric series, meaning that more recent gradients are given larger weights, and older gradients contribute less.

Conclusion: Expression for the Coefficients
The weight assigned to the gradient at iteration k is:

$$w_k = (1 - \beta)\beta^{t-k}$$

These weights sum to 1 in the infinite limit, ensuring that the momentum term remains a properly scaled moving average of past gradients.