CS/CS 316/365 Deep Learning

Activity 5

October 2, 2024

Gradient Descent

Activity needs to be handwritten. Submission will be online on canvas only.

- Show that the derivatives of the least squares loss function given below:

$$
\begin{aligned}
L[\phi] &= \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} (\mathrm{f}[x_i, \phi] - y_i)^2 \\
&= \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2
\end{aligned}
$$

are given by the expressions in this equation:

$$
\frac{\partial \ell_i}{\partial \phi} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}
$$

Solution:

$$
\begin{aligned}
\frac{\partial}{\partial \phi_0} \sum_i^I (\phi_0 + \phi_1 x_i - y_i)^2 &= \sum_i^I 2(\phi_0 + \phi_1 x_i - y_i) \cdot \frac{\partial}{\partial \phi_0}(\phi_0 + \phi_1 x_i - y_i) \\
&= \sum_i^I 2(\phi_0 + \phi_1 x_i - y_i) \cdot 1 \\
&= \sum_i^I 2(\phi_0 + \phi_1 x_i - y_i)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \phi_1} \sum_i^I (\phi_0 + \phi_1 x_i - y_i)^2 &= \sum_i^I 2(\phi_0 + \phi_1 x_i - y_i) \cdot \frac{\partial}{\partial \phi_1}(\phi_0 + \phi_1 x_i - y_i) \\
&= \sum_i^I 2(\phi_0 + \phi_1 x_i - y_i) \cdot x_i \\
&= \sum_i^I 2x_i(\phi_0 + \phi_1 x_i - y_i)
\end{aligned}
$$

- The logistic regression model uses a linear function to assign an input x to one of two classes y $\in \{0, 1\}$. For a 1D input and a 1D output, it has two parameters, ϕ_0 and ϕ_1, and is defined by:

$$Pr(y = 1|x) = \text{sig}[\phi_0 + \phi_1 x].$$

  - Plot y against x for this model for different values of ϕ_0 and ϕ_1 and explain the qualitative meaning of each parameter.
  - What is a suitable loss function for this model?
  - Compute the derivatives of this loss function with respect to the parameters.

Solution

  - The result looks like a sigmoid function which shifts to the left as we increase ϕ_0 and gets steeper as we increase ϕ_1.
  - The binary cross entropy loss:

$$L[\phi] = \sum_{i=1}^{I} -(1 - y_i) \log\Big[1 - \text{sig}[\phi_0 + \phi_1 x]\Big] - y_i \log\Big[\text{sig}[\phi_0 + \phi_1 x]\Big]$$

  - The derivatives of the sigmoid function is:

$$\frac{\partial \text{sig}[z]}{\partial z} = \frac{\exp[-z]}{(1 + \exp[-z])^2}$$

  It follows that the derivatives of the loss function are:

$$\frac{\partial L}{\partial \phi_0} = \sum_{i=1}^{I} \left(\frac{1 - y_i}{1 - \text{sig}\left[\phi_0 + \phi_1 x_i\right]} - \frac{y_i}{\text{sig}\left[\phi_0 + \phi_1 x_i\right]} \right) \frac{\exp\left[-\phi_0 - \phi_1 x_i\right]}{\left(1 + \exp\left[-\phi_0 - \phi_1 x_i\right)^2\right.}$$

$$\frac{\partial L}{\partial \phi_1} = \sum_{i=1}^{I} \left(\frac{1 - y_i}{1 - \text{sig}\left[\phi_0 + \phi_1 x_i\right]} - \frac{y_i}{\text{sig}\left[\phi_0 + \phi_1 x_i\right]} \right) \frac{x_i \cdot \exp\left[-\phi_0 - \phi_1 x_i\right]}{\left(1 + \exp\left[-\phi_0 - \phi_1 x_i\right)^2\right.}$$

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

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

Solution:

The momentum is given by:

$$\text{m}_{t+1} \leftarrow \beta \cdot \ \text{m}_t + (1 - \beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i\left[\phi_t\right]}{\partial \phi}$$

so we have:

$$
\begin{aligned}
\mathbf{m}_1 &= \beta \cdot \mathbf{m}_0 + (1-\beta) \sum_{i \in \mathcal{B}_1} \frac{\partial \ell_i [\phi_t]}{\partial \phi} \\
\mathbf{m}_2 &= \beta \cdot \mathbf{m}_1 + (1-\beta) \sum_{i \in \mathcal{B}_2} \frac{\partial \ell_i [\phi_t]}{\partial \phi} \\
&= \beta^2 \cdot \mathbf{m}_0 + \beta(1-\beta) \sum_{i \in \mathcal{B}_1} \frac{\partial \ell_i [\phi_t]}{\partial \phi} + (1-\beta) \sum_{i \in \mathcal{B}_2} \frac{\partial \ell_i [\phi_t]}{\partial \phi} \\
\mathbf{m}_3 &= \beta \cdot \mathbf{m}_2 + (1-\beta) \sum_{i \in \mathcal{B}_3} \frac{\partial \ell_i [\phi_t]}{\partial \phi} \\
&= \beta^3 \cdot \mathbf{m}_0 + \beta^2(1-\beta) \sum_{i \in \mathcal{B}_1} \frac{\partial \ell_i [\phi_t]}{\partial \phi} + \beta(1-\beta) \sum_{i \in \mathcal{B}_2} \frac{\partial \ell_i [\phi_t]}{\partial \phi} + (1-\beta) \sum_{i \in \mathcal{B}_3} \frac{\partial \ell_i [\phi_t]}{\partial \phi}
\end{aligned}
$$

and continuing in this way we see that

$$
\mathbf{m}_{t+1} = \beta^t \cdot \mathbf{m}_0 + \sum_{t'=1}^{t} \beta^{t-t'} (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i [\phi_t]}{\partial \phi}
$$

- What dimensions will the Hessian (Hessian is a square matrix with number of rows and columns equal to the number of parameters in neural network) have if the model has one million parameters?

  Solution:
  It will be 100000×1000000. This contains a trillion elements and it's impractical to invert it, or compute the eigenvalues.