SVM Loss and Derivative - Notes

Vlad Timu

September 2022

1 Introduction

The purpose of this document is to create a simple experiment to fully understand the inner working of the SVM loss function and the process of computing its partial derivatives with respect to the model parameters.

2 Problem Statement

Consider a training dataset of the form

$$X_{train} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

where each element is one example from the dataset represented as

$$x_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4}] \rightarrow X_{train} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \end{bmatrix},$$

Each sample is classified according to its row index (i.e. Sample on the row 1 has class 1. The same is true for the second and third sample) The weight matrix of the model of the form

$$W = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}$$

where each element is one set of weights, bound to one class, for all the elements of a given training sample

$$w_j = \begin{bmatrix} w_{j1} \\ w_{j2} \\ w_{j3} \\ w_{j4} \end{bmatrix} \rightarrow W = \begin{bmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \\ w_{14} & w_{24} & w_{34} \end{bmatrix},$$

3 Loss Function

3.1 Loss across training samples

The SVM loss function is generally defined as

$$L = \frac{1}{N} \sum_{i=1}^{n} L_i \tag{1}$$

where:

- N - Number of samples in the dataset
- i - Index of a single sample

3.2 Loss for one Training Sample

Note: This version of the loss does not contain the regularization term The loss component for each of the individual training samples is of the form

$$L_i = \sum_{\substack{j=1, \\ j \neq y_i}}^{C} \max(0, x_i * w_j - x_i * w_{yi} + \Delta) \tag{2}$$

where:

- C - Number of classes
- j - Index of a single class
- i - Index of a single training sample
- y_i - Index of the correct class for the i_{th} sample
- x_i - Row vector representing the i_{th} sample from the X_{train} matrix
- w_j - Column vector representing the set of weights for the j_{th} class
- w_{yi} - Column vector representing the set of weights for the correct class of the i_{th} sample
- Δ - Margin coefficient (usually set to 1)

3.3 Loss for one Training Sample Expanded

Considering the training set and the weight matrix discussed in Section 1 the loss computed for the first training sample is as follows

$$
\begin{aligned}
L_1 = \sum_{\substack{j=1,\\ j\neq 1}}^{C} \max(0, x_1 * w_j - x_1 * w_1 + \Delta) = \\
= \max(0, \overbrace{(x_{11} * w_{21} + x_{12} * w_{22} + x_{13} * w_{23} + x_{14} * w_{24})}^{x_1 * w_2 \text{ vector multiplication}} - \\
- \overbrace{(x_{11} * w_{11} + x_{12} * w_{12} + x_{13} * w_{13} + x_{14} * w_{14})}^{x_1 * w_1 \text{ - vector multiplication}} + \Delta) + \\
+ \max(0, \overbrace{(x_{11} * w_{31} + x_{12} * w_{32} + x_{13} * w_{33} + x_{14} * w_{34})}^{x_1 * w_3 \text{ - vector multiplication}} - \\
- \overbrace{(x_{11} * w_{11} + x_{12} * w_{12} + x_{13} * w_{13} + x_{14} * w_{14})}^{x_1 * w_1 \text{ - vector multiplication}} + \Delta)
\end{aligned}
\tag{3}
$$

3.4 Loss across Training Samples Expanded

Using the formula computed for the loss of one training sample, the full loss can be computed across all training samples as follows

$$
\begin{aligned}
L = \frac{1}{N} \sum_{i=1}^{N} L_i = \frac{1}{N} * (L_1 + L_2 + L_3) = \\
= \frac{1}{N} \Bigg(\overbrace{\sum_{j=1\text{j}\neq 1}^{C} \max(0, x_1 * w_j - x_1 * w_1 + \Delta)}^{L_1 \text{ - Loss for the } 1^{st} \text{ sample with correct class 1}} + \\
+ \overbrace{\sum_{j=1\text{j}\neq 2}^{C} \max(0, x_2 * w_j - x_2 * w_2 + \Delta)}^{L_2 \text{ - Loss for the } 2^{nd} \text{ sample with correct class 2}} + \\
+ \overbrace{\sum_{j=1\text{j}\neq 3}^{C} \max(0, x_3 * w_j - x_3 * w_3 + \Delta)}^{L_3 \text{ - Loss for the } 3^{rd} \text{ sample with correct class 3}} \Bigg)
\end{aligned}
\tag{4}
$$

Each of the 3 loss terms in equation (4) can be further expanded according to equation (3) to obtain the complete form of the loss function for all the training samples. The result of the loss function computation will be a single number

indicating the "correctness" of the predictions made by the model with respect to all samples used for training.

4 Loss Function Gradient

To compute the gradient of the loss function one needs to decide with respect to which parameters the gradient is to be computed. For the example presented above consider the fully expanded form of the loss function obtained from replacing each sample loss in (4) with the formula in (3).

4.1 Loss Function Gradient Across Training Samples

$$\nabla_W L = \frac{1}{N} \sum_{i=1}^{N} \nabla_W L_i \tag{5}$$

The meaning of all indices are the same as in the sections above.

4.2 Loss Function Gradient for One Sample

$$\nabla_W L_i = \begin{bmatrix} \frac{\delta L_i}{\delta w_1} & \frac{\delta L_i}{\delta w_2} & \frac{\delta L_i}{\delta w_3} \end{bmatrix} = \begin{bmatrix} \frac{\delta L_i}{\delta w_{11}} & \frac{\delta L_i}{\delta w_{21}} & \frac{\delta L_i}{\delta w_{31}} \\ \frac{\delta L_i}{\delta w_{12}} & \frac{\delta L_i}{\delta w_{22}} & \frac{\delta L_i}{\delta w_{32}} \\ \frac{\delta L_i}{\delta w_{13}} & \frac{\delta L_i}{\delta w_{23}} & \frac{\delta L_i}{\delta w_{33}} \\ \frac{\delta L_i}{\delta w_{14}} & \frac{\delta L_i}{\delta w_{24}} & \frac{\delta L_i}{\delta w_{34}} \end{bmatrix} \tag{6}$$

4.3 Loss Function Gradient for One Sample Expanded

To better understand the gradient computation process we are going to expand on the computation of one term of the matrix from equation (6). Let's take for example the last term on the second row computed for sample 1 in our dataset. That is

$$\frac{\delta L_1}{\delta w_{32}} \tag{7}$$

Considering L_1 as computed in equation (3), it follows that:

$$\frac{\delta L_1}{\delta w_{32}} = \frac{\sum_{j=1,\, j\neq 1}^{3} \max(0, x_1 * w_j - x_1 * w_1 + \Delta)}{\delta w_{32}} = \frac{L_{12} + L_{13}}{\delta w_{32}} = \frac{L_{12}}{\delta w_{32}} + \frac{L_{13}}{\delta w_{32}} \tag{8}$$

Keeping in mind that the correct class for the first sample is class 1. The expanded version of the formula above, split across all the loss terms is as follows:

- L_{12} - Loss ("correctness score") for predicting class 2 as the correct class for sample 1.

$$L_{12} = \max(0, \overbrace{(x_{11} * w_{21} + x_{12} * w_{22} + x_{13} * w_{23} + x_{14} * w_{24})}^{x_1 * w_2 \text{ vector multiplication}} - \\ - \overbrace{(x_{11} * w_{11} + x_{12} * w_{12} + x_{13} * w_{13} + x_{14} * w_{14})}^{x_1 * w_1 \text{ - vector multiplication}} + \Delta) \tag{9}$$

$$\frac{L_{12}}{\delta w_{32}} = 1(x_1 * w_2 - x_1 * w_1 + \Delta > 0)0 \tag{10}$$

The derivative of L_{12} is 0 because there are no terms dependent on w_{32} in its expression. It is worth noting that the derivative will also be 0 if the condition in the indicator function is not fulfilled. This means that the derivative of L_{12} with respect to w_{32} will always be 0.

$$\frac{L_{12}}{\delta w_{32}} = 0 \tag{11}$$

- L_{13} - Loss ("correctness score") for predicting class 3 as the correct class for sample 1.

$$L_{13} = \max(0, \overbrace{(x_{11} * w_{31} + x_{12} * w_{32} + x_{13} * w_{33} + x_{14} * w_{34})}^{x_1 * w_3 \text{ - vector multiplication}} - \\ - \overbrace{(x_{11} * w_{11} + x_{12} * w_{12} + x_{13} * w_{13} + x_{14} * w_{14})}^{x_1 * w_1 \text{ - vector multiplication}} + \Delta) \tag{12}$$

$$\frac{L_{13}}{\delta w_{32}} = 1(x_1 * w_3 - x_1 * w_1 + \Delta > 0)x_{12} \tag{13}$$

The derivative of L_{13} is x_{12} because the only term dependent on w_{32} in its expression is $x_{12} * w_{32}$. It is worth noting that the derivative will be 0 if the condition in the indicator function is not fulfilled.

$$\frac{L_{12}}{\delta w_{32}} = \begin{cases} x_{12}, & \text{if } x_1 * w_3 - x_1 * w_1 + \Delta > 0 \\ 0, & \text{if } x_1 * w_3 - x_1 * w_1 + \Delta < 0 \end{cases} \tag{14}$$

Finally, considering that the indicator function is true and replacing the value of the terms $\frac{L_{12}}{\delta w_{32}}$ and $\frac{L_{13}}{\delta w_{32}}$ in equation (8) we obtain:

$$\frac{L_1}{\delta w_{32}} = 0 + x_{12} = x_{12} \tag{15}$$

Similarly, the procedure above can be applied to compute the rest of gradient terms of $\frac{\delta L_1}{\delta w_3}$ as follows:

- $$\frac{L_1}{\delta w_{31}} = x_{11} \tag{16}$$
- $$\frac{L_1}{\delta w_{33}} = x_{13} \tag{17}$$
- $$\frac{L_1}{\delta w_{34}} = x_{14} \tag{18}$$

By replacing the elements of $\frac{\delta L_1}{\delta w_3}$ in (6) we get:

$$\nabla_W L_1 = \begin{bmatrix} \frac{\delta L_1}{\delta w_1} & \frac{\delta L_1}{\delta w_2} & \frac{\delta L_1}{\delta w_3} \end{bmatrix} = \begin{bmatrix} \frac{\delta L_1}{\delta w_{11}} & \frac{\delta L_1}{\delta w_{21}} & x_{11} \\ \frac{\delta L_1}{\delta w_{12}} & \frac{\delta L_1}{\delta w_{22}} & x_{12} \\ \frac{\delta L_1}{\delta w_{13}} & \frac{\delta L_1}{\delta w_{23}} & x_{13} \\ \frac{\delta L_1}{\delta w_{14}} & \frac{\delta L_1}{\delta w_{24}} & x_{14} \end{bmatrix} \tag{19}$$

Similarly we can compute the terms of the vectors $\frac{\delta L_1}{\delta w_1}$ and $\frac{\delta L_1}{\delta w_2}$. When replacing the elements of these column vectors in equation (19) we obtain

$$\nabla_W L_1 = \begin{bmatrix} \frac{\delta L_1}{\delta w_1} & \frac{\delta L_1}{\delta w_2} & \frac{\delta L_1}{\delta w_3} \end{bmatrix} = \begin{bmatrix} -2 * x_{11} & x_{11} & x_{11} \\ -2 * x_{12} & x_{12} & x_{12} \\ -2 * x_{13} & x_{13} & x_{13} \\ -2 * x_{14} & x_{14} & x_{14} \end{bmatrix} \tag{20}$$

The same approach as the one presented above can be used to compute the terms $\nabla_w L_2$ and $\nabla_w L_3$

$$\nabla_W L_2 = \begin{bmatrix} \frac{\delta L_2}{\delta w_1} & \frac{\delta L_2}{\delta w_2} & \frac{\delta L_2}{\delta w_3} \end{bmatrix} = \begin{bmatrix} x_{11} & -2 * x_{11} & x_{11} \\ x_{12} & -2 * x_{12} & x_{12} \\ x_{13} & -2 * x_{13} & x_{13} \\ x_{14} & -2 * x_{14} & x_{14} \end{bmatrix} \tag{21}$$

$$\nabla_W L_3 = \begin{bmatrix} \frac{\delta L_3}{\delta w_1} & \frac{\delta L_3}{\delta w_2} & \frac{\delta L_3}{\delta w_3} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{11} & -2 * x_{11} \\ x_{12} & x_{12} & -2 * x_{12} \\ x_{13} & x_{13} & -2 * x_{13} \\ x_{14} & x_{14} & -2 * x_{14} \end{bmatrix} \tag{22}$$

Note that the above gradient formulas are valid only if the condition specified by the corresponding indicator function is true. Otherwise the value of the gradients will be 0.

4.4 Loss Function Gradient Across Training Samples Expanded

After computing all the elements for $\nabla_w L_1$, $\nabla_w L_2$ and $\nabla_w L_3$, the expression in equation (5) is employed to compute an element-wise mean of the 3 matrices. The resulting matrix is then considered the final gradient that is used to update the model parameters.

$$\nabla_W L = \frac{1}{3} \sum_{i=1}^{N} \nabla_W L_i = \frac{1}{3} (\nabla_W L_1 + \nabla_W L_2 + \nabla_W L_3) =$$

$$= \frac{1}{3} \left[\begin{bmatrix} -2 * x_{11} & x_{11} & x_{11} \\ -2 * x_{12} & x_{12} & x_{12} \\ -2 * x_{13} & x_{13} & x_{13} \\ -2 * x_{14} & x_{14} & x_{14} \end{bmatrix} + \begin{bmatrix} x_{11} & -2 * x_{11} & x_{11} \\ x_{12} & -2 * x_{12} & x_{12} \\ x_{13} & -2 * x_{13} & x_{13} \\ x_{14} & -2 * x_{14} & x_{14} \end{bmatrix} + \begin{bmatrix} x_{11} & x_{11} & -2 * x_{11} \\ x_{12} & x_{12} & -2 * x_{12} \\ x_{13} & x_{13} & -2 * x_{13} \\ x_{14} & x_{14} & -2 * x_{14} \end{bmatrix} \right] \tag{23}$$