Support Vector Machines (Lab4)

Defining a loss function for MultiClass SVM:

Let $x \in \mathbb{R}^d$ be an input data point
Let N_c denote the # of classes.
Let $W \in \mathbb{R}^{d \times N_c}$ denote the parameters of the SVM.
Let N denote the total no. of data samples.
Let x_{ij} denote the j^{th} feature of the i^{th} instance.

- Given an input data point, the SVM outputs a score (s) for each class.

$$s = f(x_i) = W^T x_i \qquad \in \mathbb{R}^{N_c}$$

- The SVM loss is set up so that the SVM wants correct class for each input datapoint to have a score higher than the incorrect classes by some fixed margin (Δ). For an instance x_i, the loss

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$$

where y_i denotes the target class of x_i (label)
s_j denotes the score assigned to the j^{th} class
s_{y_i} denotes the score assigned to the correct class

(Note that this loss is always non-negative and is minimized (=0) if $s_{y_i} > s_j + \Delta$, i.e the correct class score is greater than other class scores by a margin Δ)

- If we were to simply use $\frac{1}{N} \sum_{i=1}^{N} L_i$ as the loss function, an issue can appear.
  - → Suppose W^* results in the scores for the correct class being higher than the incorrect classes by Δ.
  - → Any multiple (>1) of W^* will also satisfy the same i.e not unique!

To remove this ambiguity, we add a regularization term over W that discourages large weights: (L2 Norm)

$$R(W) \quad \frac{1}{2}\|W\|^2 = \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{N_c} W_{ij}^2$$

∴ The overall SVM loss we will use is:

$$L = \frac{1}{N} \sum_i L_i + \lambda R(W)$$

; where λ is a regularisation strength hyperparameter.

Gradient Computation:

- $\max(\cdot)$ is not a differentiable operator.
  But $\max(0, s_j - s_{y_i} + \Delta)$ will be zero only if $s_{y_i} \geq s_j + \Delta$, in which case the constraint is satisfied.
  Thus we only need the (sub-)gradient when $s_{y_i} < s_j + \Delta$

For an instance x_i

$$\nabla_W L_i = \begin{bmatrix} \frac{\partial L_i}{\partial W_{11}} & \frac{\partial L_i}{\partial W_{12}} & \cdots & \frac{\partial L_i}{\partial W_{1N_c}} \\ \vdots & & & \\ \frac{\partial L_i}{\partial W_{d1}} & \frac{\partial L_i}{\partial W_{d2}} & \cdots & \frac{\partial L_i}{\partial W_{dN_c}} \end{bmatrix} \in \mathbb{R}^{d \times N_c}$$

We have,

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$$

$$= \sum_{j \neq y_i} \max(0, (W^T x_i)_j - (W^T x_i)_{y_i} + \Delta)$$

$$= \max(0, (W_{11} x_{i1} + W_{21} x_{i2} + \ldots W_{d1} x_{id}) - (W_{1y_i} x_{i1} + W_{2y_i} x_{i2} + \ldots W_{dy_i} x_{id}) + \Delta)$$

$$+ \max(0, (W_{12} x_{i1} + W_{22} x_{i2} + \ldots W_{d2} x_{id}) - (W_{1y_i} x_{i1} + W_{2y_i} x_{i2} + \ldots W_{dy_i} x_{id}) + \Delta)$$

$$+ \vdots$$

$$\max(0, (W_{1N_c} x_{i1} + W_{2N_c} x_{i2} + \ldots W_{dN_c} x_{id}) - (W_{1y_i} x_{i1} + W_{2y_i} x_{i2} + \ldots W_{dy_i} x_{id}) + \Delta)$$

$\therefore$ If $(W^T x_i)_j - (W^T x_i)_{y_i} \geq \Delta$: (else gradient is zero).

For $j \neq y_i$:

$$\frac{\partial L_i}{\partial W_{1j}} = x_{i1}, \quad \vdots \quad \frac{\partial L_i}{\partial W_{dj}} = x_{id}$$

$$\therefore \boxed{\frac{\partial L_i}{\partial W_{:,j}} = x_i} \quad (\text{for } j \neq y_i)$$

for $j = y_i$,

$$\frac{\partial L_i}{\partial W_{1y_i}} = -k \cdot x_{i1}, \quad \vdots \quad \frac{\partial L_i}{\partial W_{dy_i}} = -k \cdot x_{id}$$

$$\therefore \boxed{\frac{\partial L_i}{\partial W_{:,j}} = -k\, x_i} \quad (\text{for } j = y_i)$$

where k = no. of classes satisfying $s_{y_i} > s_j + \Delta$

i.e. $k = \sum_j \mathbb{1}(s_{y_i} > s_j + \Delta)$

$$\left(\mathbb{1}(x) = \begin{cases} 1 & ; x = \text{True} \\ 0 & ; x = \text{False} \end{cases} \right)$$

- Gradient w.r.t $R(W)$.

since $R(W) = \frac{1}{2} \|W\|^2$

$$= \frac{1}{2} \sum_i \sum_j W_{ij}^2$$

$$\frac{\partial R(W)}{W_{ij}} = W_{ij}$$

Using the above, obtain the gradient for the total loss.

$$\nabla_W L = \nabla_W \left(\frac{1}{N} \sum_i L_i + \lambda R(W) \right)$$

$$= \frac{1}{N} \sum_i \nabla_W L_i + \lambda \nabla_W R(W)$$