MACHINE LEARNING FUNDAMENTALS - LS2026
SEMINAR: DEEP LEARNING

CZECH TECHNICAL UNIVERSITY IN PRAGUE
V. FRANC

Assignment 1. Learning a linear Support Vector Machine classifier $h(\mathbf{x}; \mathbf{w}, b) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ from a training set $T_m = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{+1, -1\} \mid i = 1, \ldots, m\}$ leads to the following convex quadratic program:

$$(\hat{\mathbf{w}}, \hat{b}, \hat{\boldsymbol{\xi}}) = \arg \min_{\mathbf{w} \in \mathbb{R}^d,\, b \in \mathbb{R},\, \boldsymbol{\xi} \in \mathbb{R}^m} \left[\frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \right] \tag{1}$$

subject to

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad i = 1, \ldots, m.$$

a) Derive a stochastic gradient descent algorithm for solving the primal soft-margin SVM optimization problem in (1).
b) Compare the resulting algorithm with the Perceptron learning algorithm.

Solution 1. Note that for any fixed $(\mathbf{w}, b)$, the optimal slack variable reads

$$\xi_i^\star = \max\{0,\, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}.$$

Hence the primal problem is equivalent to the unconstrained problem

$$\min_{\mathbf{w}, b} F(\mathbf{w}, b) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^m \max\{0,\, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}.$$

Let

$$\ell_i(\mathbf{w}, b) = \max\{0,\, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}.$$

For a randomly chosen index $i_t \sim \text{Uniform}\{1, \ldots, m\}$, define

$$s_t = y_{i_t}(\mathbf{w}_t^\top \mathbf{x}_{i_t} + b_t).$$

A stochastic subgradient of F is

$$\mathbf{g}_{\mathbf{w}}^{(t)} = \lambda \mathbf{w}_t - y_{i_t} \mathbf{x}_{i_t} [\![s_t < 1]\!], \qquad g_b^{(t)} = -y_{i_t} [\![s_t < 1]\!].$$

Therefore SGD uses

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_{\mathbf{w}}^{(t)}, \qquad b_{t+1} = b_t - \eta_t g_b^{(t)}.$$

Equivalently,

$$\boxed{\begin{aligned} \text{if } s_t < 1: \quad & \mathbf{w}_{t+1} = (1 - \eta_t \lambda) \mathbf{w}_t + \eta_t y_{i_t} \mathbf{x}_{i_t}, \\ & b_{t+1} = b_t + \eta_t y_{i_t}, \\ \text{if } s_t \geq 1: \quad & \mathbf{w}_{t+1} = (1 - \eta_t \lambda) \mathbf{w}_t, \\ & b_{t+1} = b_t. \end{aligned}}$$

Algorithm: SGD for primal soft-margin SVM

Initialize $\mathbf{w}_0 = \mathbf{0},\ b_0 = 0.$
For $t = 0, 1, \ldots, T-1:$
  sample $i_t \sim \mathrm{Uniform}\{1, \ldots, m\};$
  $s_t \leftarrow y_{i_t}(\mathbf{w}_t^\top \mathbf{x}_{i_t} + b_t);$
  if $s_t < 1:$
    $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t\lambda)\mathbf{w}_t + \eta_t y_{i_t}\mathbf{x}_{i_t};$
    $b_{t+1} \leftarrow b_t + \eta_t y_{i_t};$
  else:
    $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t\lambda)\mathbf{w}_t;$
    $b_{t+1} \leftarrow b_t.$
Return $(\mathbf{w}_T, b_T)$ or averaged iterates.