Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2022

# Writing Assignment 3

**Issued:** Saturday 29th October, 2022 **Due:** Tuesday 8th November, 2022

## POLICIES

- **Acknowledgments:** We expect you to make an honest effort to solve the problems individually. As we sometimes reuse problem set questions from previous years, covered by papers and web pages, we expect the students **NOT** to copy, refer to, or look at the solutions in preparing their answers (relating to an unauthorized material is considered a violation of the honor principle). Similarly, we expect you to not google directly for answers (though you are free to google for knowledge about the topic). If you do happen to use other material, it must be acknowledged here, with a citation on the submitted solution.
- **Required homework submission format:** You can submit homework either as one single PDF document or as handwritten papers. Written homework needs to be provided during the class on the due date, and a PDF document needs to be submitted through Tsinghua's Web Learning (`http://learn.tsinghua.edu.cn/`) before the end of the due date.
- **Collaborators:** In a separate section (before your answers), list the names of all people you collaborated with and for which question(s). If you did the HW entirely on your own, **PLEASE STATE THIS**. Each student must understand, write, and hand in answers of their own.

3.1. (Kernel SVM) Suppose we are given a training dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$ consisting of $m$ independent examples, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^n$ is $n$-dimension vector, and $y^{(i)} \in \{-1, +1\}$. When the data are not linearly separable, consider the Kernel-SVM given by

$$\begin{aligned} \underset{\boldsymbol{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b) \geq 1, \quad i = 1, \ldots, m, \end{aligned} \tag{1}$$

where $\phi(\boldsymbol{x})$ is a mapping function $\phi(\boldsymbol{x}) : (x_1, x_2) \mapsto \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$.

(a) (1 point) Prove that $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{\text{def}}{=} \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$ is positive semi-definite symmetric, i.e. for any vector $\boldsymbol{v} \in \mathbb{R}^m$, $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{v} \geq 0$ .

(b) (2 points) Given data set $\left\{((1, \sqrt{2})^{\mathrm{T}}, 1), ((\sqrt{2}, 1)^{\mathrm{T}}, 1), ((2, \sqrt{2})^{\mathrm{T}}, -1)\right\}$, derive the optimal value of $\boldsymbol{w}^*$ and $b^*$ in (1).

(c) (1 point) In (b), for new sample $(4\sqrt{2}, 1)^{\mathrm{T}}$, make your decision of classification.

3.2. (Least-Squares SVM) Suppose we are given a training dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$ consisting of $m$ independent examples, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^n$ is $n$-dimension vector, and $y^{(i)} \in \{-1, 1\}$. The Least-Squares Support Vector Machine (LS-SVM) aims to construct a linear model $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b$ in a given feature space, i.e. $\phi(\boldsymbol{x}) : \mathcal{X} \to \mathbb{F}$, that is able to distinguish between examples drawn from different categories $\mathcal{C}^-$ and $\mathcal{C}^+$, such that

$$\boldsymbol{x} \in \begin{cases} \mathcal{C}^+, & f(\boldsymbol{x}) \geq 0 \\ \mathcal{C}^-, & o.w. \end{cases} .$$

The optimal model parameters $(\boldsymbol{w}^*, b^*)$ are given by solving a constrained optimization problem,

$$\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{2\mu}\sum_{i=1}^{m}\epsilon_i^2 \\ \text{subject to} & y_i = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b + \epsilon_i, \quad i = 1, \ldots, m, \end{array} , \tag{2}$$

where $\mu$ is a regularization hyper-parameter. The primal Lagrangian for this optimization problem (2) gives the unconstrained minimization problem,

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{2\mu}\sum_{i=1}^{m}\epsilon_i^2 - \sum_{i=1}^{m}\alpha_i[\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b + \epsilon_i - y_i], \tag{3}$$

where $\boldsymbol{\alpha} \overset{\text{def}}{=} [\alpha_1, \ldots \alpha_m]^T$ is a vector of Lagrange multipliers.

(a) (2 points) Give the KKT optimality conditions for this problem. (Hint: Set $\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \epsilon_i} = \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0$)

(b) (2 points) Denoting that $\boldsymbol{K}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \overset{\text{def}}{=} \left\langle \phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)}) \right\rangle$, prove that

$$\begin{bmatrix} \boldsymbol{K} + \mu\boldsymbol{I} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^* \\ b^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}.$$

3.3. (Back Propagation)(2 points) Consider the backpropagation on the hidden layer in a neural network. Given a batch of input feature $X = [x^{(1)}, x^{(2)}, \cdots, x^{(M)}]^T$ (Shape: $M \times D_0$), a set of weight $\{W \in \mathbb{R}^{D_0 \times D_1},\ b \in \mathbb{R}^{D_1 \times 1}\}$, and element-wise Sigmoid activation function $\sigma(\cdot)$. The forward propagation on this hidden layer is given by:

$$F_1 = XW + \mathbb{1}_M b^T, \quad F_2 = \sigma(F_1)$$

where $\mathbb{1}_M$ is a vector composed of 1 in length $M$. $\sigma(\cdot)$ is element-wise Sigmoid function:

$$[\sigma(X)]_{ij} = \frac{1}{1 + \exp(-X_{ij})}$$

Then in the backpropagation stage, suppose we already know the gradients for some scalar loss function $l$ with respect to $F_2$ as $\nabla_{F_2} l$. Proceed the backpropagation and show that $\nabla_{F_1} l = (\nabla_{F_2} l) \odot F_2 \odot (1 - F_2)$. where $(A \odot B)_{ij} = A_{ij}B_{ij}$ is element-wise production.