Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2021

## Writing Assignment 2

**Issued:** Friday 22nd October, 2021 **Due:** Wednesday 3rd November, 2021

2.1. (Kernel Regression Least Square) Suppose we are given a dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^m$ consisting of $m$ independent samples, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^n$ is a $n$-dimension vector, and $y^{(i)} \in \mathbb{R}$. Now, we aim to learn a linear model $f(\boldsymbol{x}) = \boldsymbol{\theta}^{\mathrm{T}}\phi(\boldsymbol{x})$ in a given feature space, i.e. $\phi(\boldsymbol{x}) : \mathcal{X} \to \mathcal{F}$, with regularization term $\lambda\|\boldsymbol{\theta}\|_2^2$. The loss function of the linear regression problem can be given as,

$$\sum_{i=1}^{m}[y^{(i)} - (\boldsymbol{\theta}^{\mathrm{T}}\phi(\boldsymbol{x}^{(i)}))]^2 + \lambda\|\boldsymbol{\theta}\|_2^2. \tag{1}$$

(a) (2 points) Prove that the optimal parameter $\boldsymbol{\theta}^*$ is in the span of features $\{\phi(\boldsymbol{x}^{(i)})\}_{i=1}^m$, i.e. $\boldsymbol{\theta}^* = \sum_{i=1}^m c_i\phi(\boldsymbol{x}^{(i)})$, where $c_i$ then becomes the term needed to be optimized.(Hint: Set the differentiation of the loss over $\boldsymbol{\theta}$ to 0)

(b) (2 points) The mapping function $\phi(\cdot)$ can often result to a high-dimensional or infinite feature $\phi(\boldsymbol{x})$. Thus, we adopt a kernel $\boldsymbol{K}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \stackrel{\text{def}}{=} \left\langle\phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)})\right\rangle$ to make the calculation easier. Based on (a), we know that

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^{\mathrm{T}}\phi(\boldsymbol{x}) = \sum_{i=1}^{m} c_i\left\langle\phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x})\right\rangle \tag{2}$$

$$\|\boldsymbol{\theta}\|_2^2 = \left\langle\sum_{i=1}^{m} c_i\phi(\boldsymbol{x}^{(i)}), \sum_{j=1}^{m} c_j\phi(\boldsymbol{x}^{(j)})\right\rangle = \boldsymbol{c}^T\boldsymbol{K}\boldsymbol{c}, \tag{3}$$

where $\boldsymbol{c} \stackrel{\text{def}}{=} [c_1, \dots c_m]^T$ and $\boldsymbol{K} \in \mathbb{R}^{m\times m}$ with the $(i,j)$-th entry defined as $\boldsymbol{K}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. Now please rewrite the loss function (1) using $\boldsymbol{c}$ and $\boldsymbol{K}$, and give the optimal parameter $\boldsymbol{c}^*$.

2.2. (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,

$$\begin{aligned} \underset{\boldsymbol{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{2\mu}\sum_{i=1}^{m}\epsilon_i^2 \\ \text{subject to} \quad & y_i = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b + \epsilon_i, \quad i = 1, \dots, m, \end{aligned} \tag{4}$$

where $\mu$ is a regularization hyper-parameter. The primal Lagrangian for this optimisation problem (4) gives the unconstrained minimisation problem,

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{2\mu}\sum_{i=1}^{m}\epsilon_i^2 - \sum_{i=1}^{m}\alpha_i[\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b + \epsilon_i - y_i], \tag{5}$$

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

(a) (1 point) 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 $\boldsymbol{K}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \stackrel{\text{def}}{=} \langle\phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)})\rangle$, prove that

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

(c) (1 point) Let $\boldsymbol{M} \stackrel{\text{def}}{=} \boldsymbol{K} + \mu\boldsymbol{I}$, prove that

$$\boldsymbol{\alpha}^* = \boldsymbol{M}^{-1}(\boldsymbol{y} - b^*\mathbf{1}), \qquad b^* = \frac{\mathbf{1}^T\boldsymbol{M}^{-1}\boldsymbol{y}}{\mathbf{1}^T\boldsymbol{M}^{-1}\mathbf{1}},$$

where $\boldsymbol{y} \stackrel{\text{def}}{=} [y_1, \ldots y_m]^T$ and $\mathbf{1} \stackrel{\text{def}}{=} [1, \ldots 1]^T$.

2.3. (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

$$\begin{aligned} &\underset{\boldsymbol{w},b}{\text{minimize}} && \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ &\text{subject to} && y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b) \geq 1, \quad i = 1, \ldots, m, \end{aligned} \tag{6}$$

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 $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \stackrel{\text{def}}{=} \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$ 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 (6).

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