THU-70250403, Convex Optimization (Fall 2020) **Homework: 6**

Classification and Parameter Estimation

Lecturer: Li Li li-li@tsinghua.edu.cn

Student:

Twin Support Vector Machine (TWSVM) aims to classify two classes of data by finding two non-parallel hyper-planes in such a manner that the first hyper-plane locates near to the data samples of the first class while distant from the second class of data samples. Vice verse, the second hyper-plane locates near to the data samples of the second class while distant from the first class of data samples [1]-[2].[1]

Let the samples of positive and negative classes are denoted by m and n respectively and positive and negative class data samples are represented by data matrices $X_1 \in R^{k\times m}$ and $X_2 \in R^{k\times n}$. $X_1 = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1m})$ and $X_2 = (\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n})$.

Suppose two non-parallel hyper-planes in k-dimensional real space R_k are parameterized as

$$\boldsymbol{w}_1^T\boldsymbol{x} + b_1 = 0 \tag{1}$$
$$\boldsymbol{w}_2^T\boldsymbol{x} + b_2 = 0 \tag{2}$$

where, symbols $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ indicate normal vectors to the hyperplane,b_1 and b_2 are bias terms. The formulation of TWSVM for linear case is obtained.

The primal problem of soft margin Linear TWSVM can be formulated as

$$\min_{\boldsymbol{w}_1,b_1,\boldsymbol{\xi}} \quad \frac{1}{2}\sum_{i=1}^{m}\left(\boldsymbol{w}_1^T\boldsymbol{x}_{1i} + b_1\right)^2 + C_1\sum_{j=1}^{n}\xi_j \tag{3}$$
$$\text{s.t.} \quad -\left(\boldsymbol{w}_1^T\boldsymbol{x}_{2j} + b_1\right) + \xi_j \geq 1,\ j = 1, \ldots, n \tag{4}$$
$$\xi_j \geq 0,\ j = 1, \ldots, n \tag{5}$$
$$\min_{\boldsymbol{w}_2,b_2,\boldsymbol{\eta}} \quad \frac{1}{2}\sum_{j=1}^{n}\left(\boldsymbol{w}_2^T\boldsymbol{x}_{2j} + b_2\right)^2 + C_2\sum_{i=1}^{m}\eta_i \tag{6}$$
$$\text{s.t.} \quad -\left(\boldsymbol{w}_2^T\boldsymbol{x}_{1i} + b_2\right) + \eta_i \geq 1,\ i = 1, \ldots, m \tag{7}$$
$$\eta_i \geq 0,\ i = 1, \ldots, m \tag{8}$$

where slack variables and penalty parameters are represented by $\boldsymbol{\xi}$,$\boldsymbol{\eta}$ and C_1,C_2 respectively.

We can get the generalized Lagrangian function of the first subproblem as

$$L(\boldsymbol{w}_1, b_1, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{w_1}^T\boldsymbol{x_{1i}} + b_1)^2 + C_1\sum_{j=1}^{n}\xi_j + \sum_{j=1}^{n}\alpha_j((\boldsymbol{w_1}^T\boldsymbol{x_{2j}} + b_1) - \xi_j + 1) - \sum_{j=1}^{n}\beta_j\xi_j \tag{9}$$

where $\boldsymbol{\alpha} \in \mathbb{R}_+^L$ and $\boldsymbol{\beta} \in \mathbb{R}_+^L$ are the associated Lagrange multipliers.

[1] Some researchers argued that the Twin SVM cannot be viewed as a SVM, since all the data points are active to generate the solution of the optimization problems and no sparse supporting vectors exist.

Considering the Karush-Kuhn-Tucker (KKT) conditions of Eq.(9), we have

$$\frac{\partial L(\boldsymbol{w}_1, b_1, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{w}_1} = \sum_{i=1}^{m} (\boldsymbol{x}_{1i}^T \boldsymbol{w}_1 + b_1)\boldsymbol{x}_{1i} + \sum_{j=1}^{n} \alpha_j \boldsymbol{x}_{2j} = 0 \tag{10}$$

$$\frac{\partial L(\boldsymbol{w}_1, b_1, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial b_1} = \sum_{i=1}^{m} (\boldsymbol{w}_1^T \boldsymbol{x}_{1i} + b_1) + \sum_{j=1}^{n} \alpha_j = 0 \tag{11}$$

$$\frac{\partial L(\boldsymbol{w}_1, b_1, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \xi_j} = C_1 - \beta_j - \alpha_j = 0 \tag{12}$$

$$\alpha_j \left[(\boldsymbol{w}_1^T \boldsymbol{x}_{2j} + b_1) - \xi_j + 1 \right] = 0, \quad \beta_j \xi_j = 0, \quad j = 1, \ldots, n \tag{13}$$

Summarizing all these up, we have

$$\begin{bmatrix} X_1 \\ \mathbf{1}^T \end{bmatrix} \begin{bmatrix} X_1^T & \mathbf{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{b}_1 \end{bmatrix} + \begin{bmatrix} X_2 \\ \mathbf{1}^T \end{bmatrix} \boldsymbol{\alpha} = 0 \tag{14}$$

Let $A = \begin{bmatrix} X_1^T & \mathbf{1} \end{bmatrix}$, $X_1 = [\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1m}]$ and $B = \begin{bmatrix} X_2^T & \mathbf{1} \end{bmatrix}$, $X_2 = [\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n}]$, $\boldsymbol{u}_1 = [\boldsymbol{w}_1 \ \boldsymbol{b}_1]^T$.

We can rewrite Eq.(14) as

$$A^T A \boldsymbol{u}_1 + B^T \boldsymbol{\alpha} = 0 \tag{15}$$

So, we get

$$\boldsymbol{u}_1 = -\left(A^T A\right)^{-1} B^T \boldsymbol{\alpha} \tag{16}$$

Substituting the above equations to the Lagrange dual function, we obtain the dual problem as

$$\max_{\boldsymbol{\alpha}} \quad \mathbf{1}^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T B (A^T A)^{-1} B^T \boldsymbol{\alpha} \tag{17}$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C_1, \quad i = 1, \ldots, n \tag{18}$$

In the same manner, normal vector and bias for the second class are achieved by solving below equation:

$$\boldsymbol{u}_2 = -\left(B^T B\right)^{-1} A^T \boldsymbol{\gamma} \tag{19}$$

where $\boldsymbol{\gamma} \in \mathbb{R}_+^L$ and $\boldsymbol{\zeta} \in \mathbb{R}_+^L$ are the associated Lagrange multipliers for the generalized Lagrangian function of the first subproblem.

And the dual problem for the second class is wrriten as

$$\max_{\boldsymbol{\gamma}} \quad \mathbf{1}^T \boldsymbol{\gamma} - \frac{1}{2} \boldsymbol{\gamma}^T A (B^T B)^{-1} A^T \boldsymbol{\gamma} \tag{20}$$

$$\text{s.t.} \quad 0 \leq \gamma_i \leq C_2, \quad i = 1, \ldots, m \tag{21}$$

When all the parameters are known, TWSVM determines whether a new data sample $\boldsymbol{z}$ is assigned to a class i by using following decision function:

$$Class\ i = \min_i \left| \boldsymbol{w}_i^T \boldsymbol{z} + b_i \right| \ , \ i = 1, 2 \tag{22}$$

That is, the perpendicular distance of the test data sample is calculated from each hyper-plane and pattern is assigned to the class from which its distance is lesser.

(Please choose either one problem from the first two problems and finish it together with Problem 3.)

Problem 1

Please write a program code snippet (a CVX based matalb program code is preferred) to implement both primary and dual solution of the Twin SVM. Test your code with some data and show the outcomes.

Problem 2

Please extend the above Twin SVM model for binary classification problem for multi-class problems.

Problem 3

This problem appears in the 2013 Final Exam of EE364a: Convex Optimization I, Stanford University.

Maximum likelihood estimation for an affinely transformed distribution. Let $\boldsymbol{z}$ be a random variable on $\mathbb{R}^n$ with density $p_{\boldsymbol{z}}(\boldsymbol{u}) = \exp\{-\phi(|\boldsymbol{u}|_2)\}$, where $\phi : \mathbb{R} \to \mathbb{R}$ is convex and increasing. Examples of such distributions include the standard normal $N(0, \sigma^2 I)$, with $\phi(u) = (u)_+^2 + \alpha$, and the multivariable Laplacian distribution, with $\phi(u) = (u)_+ + \beta$, where α and β are normalizing constants, and $(a)_+ = \max\{a, 0\}$.

Now let $\boldsymbol{x}$ be the random variable $\boldsymbol{x} = A\boldsymbol{z} + \boldsymbol{b}$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular. The distribution of $\boldsymbol{x}$ is parametrized by A and $\boldsymbol{b}$.

Suppose $\boldsymbol{x}_1, \ldots, x_N$ are independent samples from the distribution of $\boldsymbol{x}$. Explain how to find a maximum likelihood estimate of A and $\boldsymbol{b}$ using convex optimization. If you make any further assumptions about A and $\boldsymbol{b}$ (beyond invertiblility of A), you must justify it. Hint: The density of $\boldsymbol{x} = A\boldsymbol{z} + \boldsymbol{b}$ is given by $p_{\boldsymbol{x}}(\boldsymbol{v}) = \frac{1}{|\det(A)|} p_{\boldsymbol{z}}\left(A^{-1}(\boldsymbol{v} - \boldsymbol{b})\right)$.

References

[1] Jayadeva, R. Khemchandani, S. Chandra, "Twin support vector machine for pattern classification," *IEEE Tractions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905-910, 2007.

[2] D. Tomar, S. Agarwal, "Twin Support Vector Machine: A review from 2007 to 2014," *Egyptian Informatics Journal*, vol. 16, pp. 55-69, 2015.