SUPPORT VECTOR MACHINES & NEURAL NETWORKS

LECTURE 6 – SUPPORT VECTOR MACHINES PART # III

A. Bi-classification

History, LSVM, Approximate LSVM, Soft LSVM, Kernel-based linear SVM, nonlinear SVM

B. Multi-classification

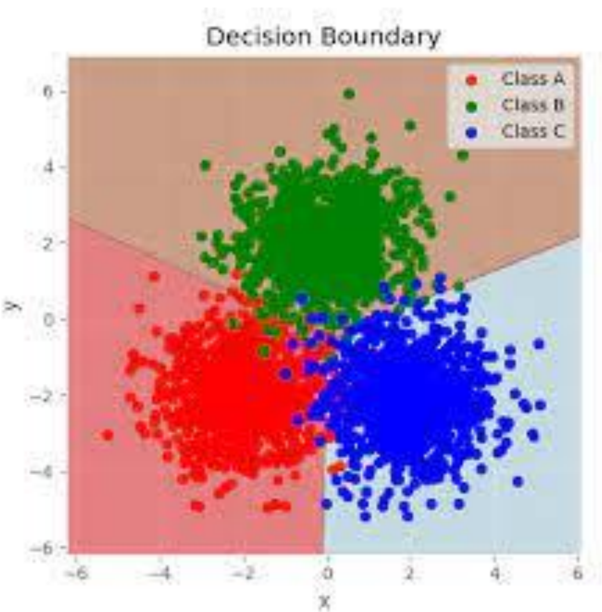
OVO, OVA, Twin SVM

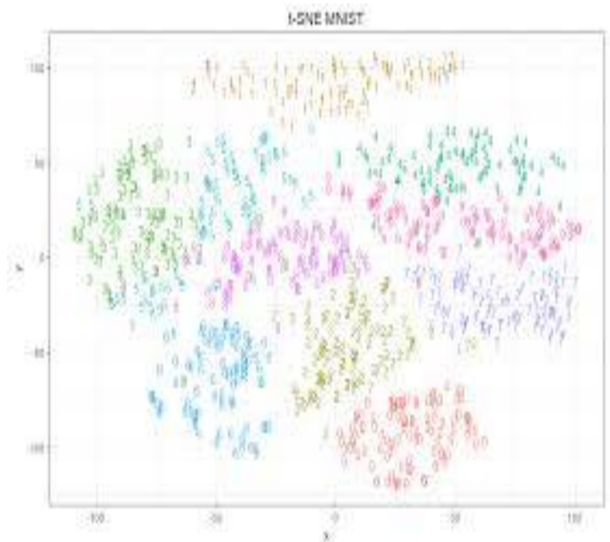
C. Prediction

Support Vector Regression (SVR)

Multi-class classification

- Multi-class classification (multi-classification) is a problem of classifying instances into one of three or more classes.



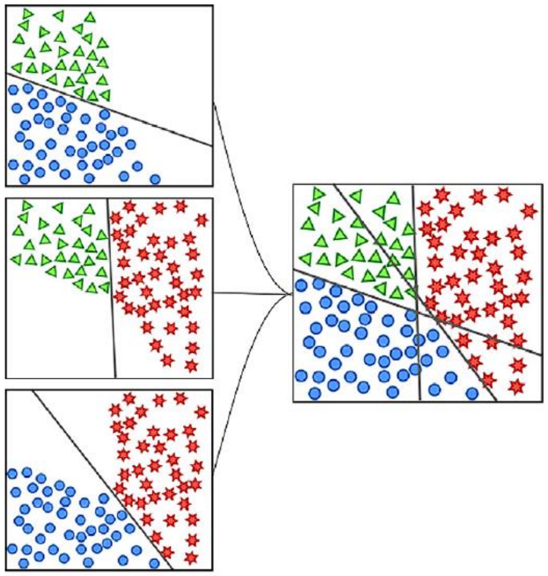


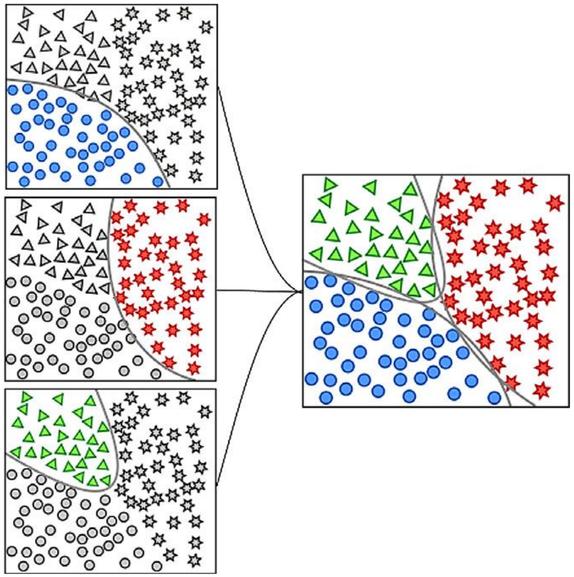
- Pictures taken from Wikipedia

Basic ideas of multiclass classification

- If we are given a dataset in C classes and a new data-point $\boldsymbol{x} \in \mathbb{R}^n$, we may consider two commonly adopted approaches to determine which class $\boldsymbol{x}$ belongs to:

1. OVO (one vs. one) approach
2. OVA (one vs. all) approach





Basic ideas of multiclass classification

- 1. OVO (one vs. one) approach

  - Take each pair of different classes (C_i, C_j).

  - Label all data-points in C_i with a label $+1$;

    Label all data-points in C_j $\text{with a label } -1$.

  - Apply a bi-class SVM classifier to find $(\boldsymbol{w}, b)$

    and determine $\boldsymbol{x} \in C_i \; or \; \boldsymbol{x} \in C_j$.

  - Assign a voting score to each class of the pair

    $$Score(C_i) = \begin{cases} 1, & if\ \boldsymbol{x} \in C_i \\ 0, & if\ \boldsymbol{x} \in C_j \end{cases}$$

  - Sum up scores/votes over all pairs involving C_i.

  - Decision: $\boldsymbol{x}$ belongs to the class with the hightest total score.

OVO related issues

- How many SVM involved?

  # of all possible pairs = $\mathcal{C}(\mathcal{C}-1)/2$
- Tie breaker?

  secondary score?
- Better scoring measure?

Basic ideas of multiclass classification

- 2. OVA (one vs. all of the rest) approach

  - Take each class C_i.

  - Label all data-points in C_i with a label $+1$;

  Label all data-points not in C_i $\text{with a label } -1$.

  - Apply a bi-class SVM classifier to find $(\boldsymbol{w}, b)$

  and determine $\boldsymbol{x} \in C_i\ or\ \boldsymbol{x} \notin C_i$.

  - Assign a score to each class

  $Score(C_i) = \boldsymbol{w}^T\boldsymbol{x} + b$

  - Decision: $\boldsymbol{x}$ belongs to the class with hightest score.

OVA related issues

- How many SVM involved?

    # of all possible pairs = $\mathcal{C}$ ($\ll \mathcal{C}(\mathcal{C}-1)/2$)
- Imbalanced datasets

    quality of results?
- Better scoring ?

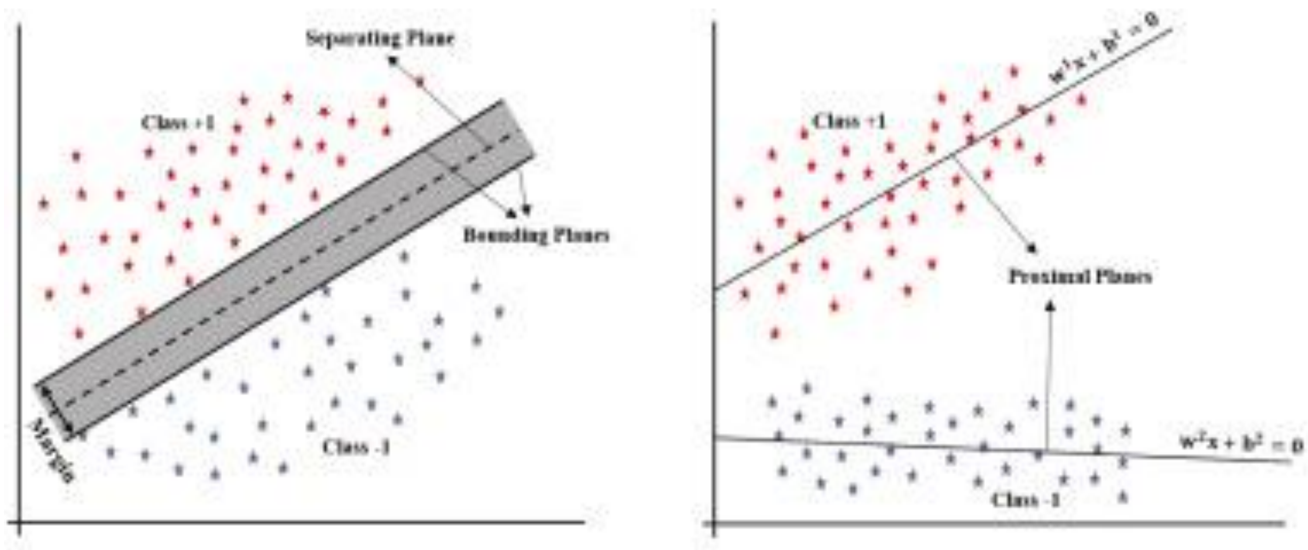
- Any approaches other than OVO and OVA ?

    - tournament ?
- Which SVM model to use?

Beyond LSSVM and KSSVM

- Motivation: Who says we should use only one separation hyperplane (surface) for bi-classification?

Picture from ScienceDirect.com



(a) (b)

Twin support vector machines - TWSVM

- Basic ideas:

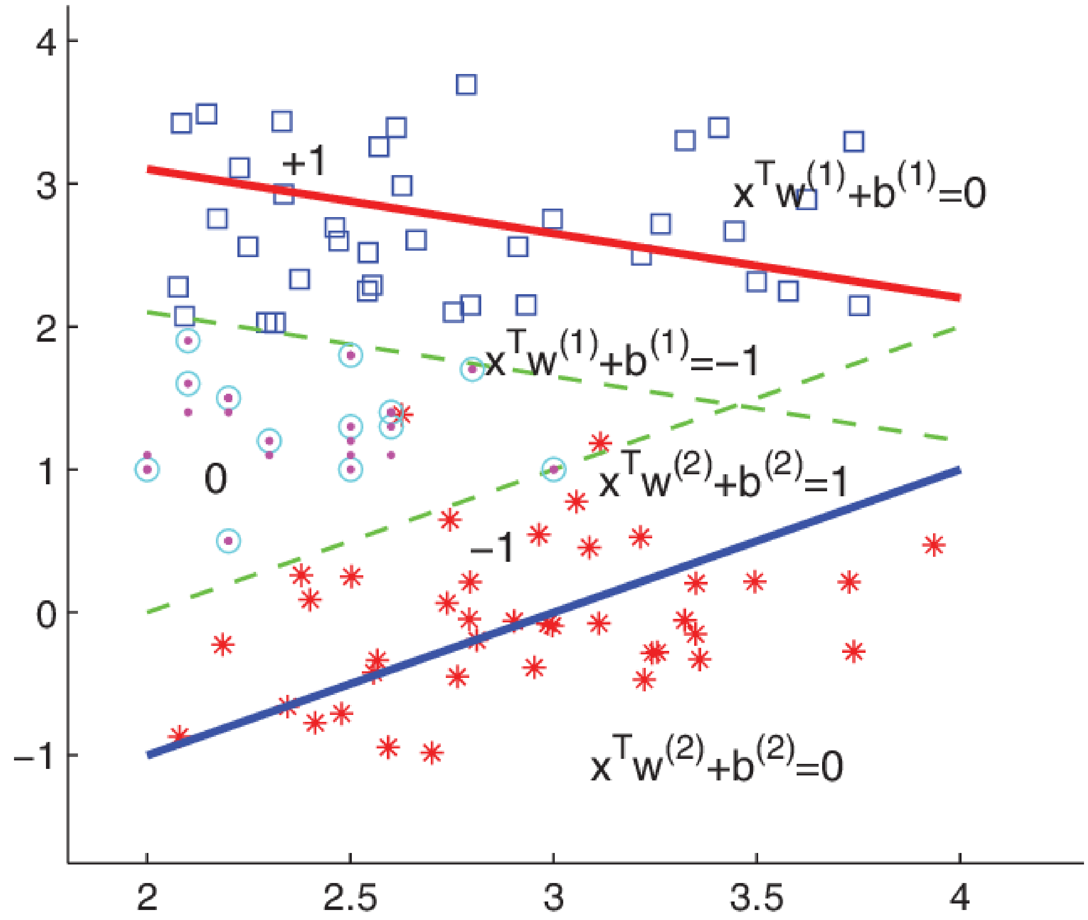
  (i) Use two non-parallel SVM to separate two distinct classes of data;

  (ii) All points in one class center around a corresponding SVM separation hyperplane while all points in the other class are kept away from this hyperplane for a safe distance;

  (iii) When a new point comes into the picture, it is classified based on its “distance” to each hyperplane.

Realization of TWSVM

- Picture from researchgate.net



Twin support vector machines

- References:

1. O.L. Mangasarian and E.W. Wild, “Multisurface Proximal Support Vector Classification via Generalized Eigenvalues,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 1, pp. 69-74, 2006.

2. Jayadeva, R. Khemchandani and S. Chandra, “Twin Support Vector Machines for Pattern Classification,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 5, pp. 905-910,. 2007.

Twin support vector machines

Problem Setting:

- Dataset $S = \{\boldsymbol{x}^i \in \mathbb{R}^n \mid i = 1,2,\dots,N\}$ of N data points of n attributes.

- Two classes: $S = S_A \cup S_B, S_A \cap S_B = \emptyset$ with $|S_A| = N_A$ and $|S_B| = N_B$.

- Denote $S_A = \{\boldsymbol{x}_A^i \in \mathbb{R}^n \mid i = 1,2,\dots,N_A\}$,

  $S_B = \{\boldsymbol{x}_B^j \in \mathbb{R}^n \mid j = 1,2,\dots,N_B\}$.

- Labels

  $y_i = +1, \text{for } \boldsymbol{x}_A^i \in S_A, i = 1,2,\dots,N_A$

  $y_j = -1, \text{for } \boldsymbol{x}_B^j \in S_B, j = 1,2,\dots,N_B$

Twin support vector machines

Problem Setting:

- Separation hyperplanes:

  For S_A Class: $\boldsymbol{w}^{(\mathbf{1})^{\boldsymbol{T}}}\boldsymbol{x} + b^{(1)} = 0$

  For S_B Class: $\boldsymbol{w}^{(\mathbf{2})^{\boldsymbol{T}}}\boldsymbol{x} + b^{(2)} = 0$

  $\boldsymbol{w}^{(i)} \in \mathbb{R}^n, b^{(i)} \in \mathbb{R}, i = 1, 2.$

  Denote $\boldsymbol{u}^{(i)} = \begin{pmatrix} \boldsymbol{w}^{(i)} \\ b^{(i)} \end{pmatrix} \in \mathbb{R}^{n+1}, i = 1, 2.$

- Data preparation

  data records of S_{A} : $X_{\mathrm{A}}^T = [x_A^1, \dots, x_A^{N_A}] \in \boldsymbol{M}^{n \times N_A}$

  data records of S_{B} : $X_{\mathrm{B}}^T = [x_B^1, \dots, x_B^{N_B}] \in \boldsymbol{M}^{n \times N_B}$

  unit vector $\boldsymbol{e}_A \in \mathbb{R}^{N_A}$: a column vector with all elements being 1.

  unit vector $\boldsymbol{e}_B \in \mathbb{R}^{N_B}$: a column vector with all elements being 1.

  Denote $\hat{X}_A^T = \begin{pmatrix} X_{\mathrm{A}}^T \\ e_A^T \end{pmatrix} \in \boldsymbol{M}^{(n+1) \times N_A}$ and $\hat{X}_B^T = \begin{pmatrix} X_{\mathrm{B}}^T \\ e_B^T \end{pmatrix} \in \boldsymbol{M}^{(n+1) \times N_B}$

Twin soft support vector machine model - TWSSVM

(TWSSVM-1)

$$Min \quad \frac{1}{2}\sum_{i=1}^{N_A}\left(\boldsymbol{w}^{(1)^T}\boldsymbol{x}_A^i + b^{(1)}\right)^2 + C_1 \sum_{j=1}^{N_B}\xi_j^{(1)}$$

$$s.t. \quad y_j\left(\boldsymbol{w}^{(1)^T}\boldsymbol{x}_B^j + b^{(1)}\right) \geq 1 - \xi_j^{(1)}, \; j = 1, \ldots, N_B$$

$$\boldsymbol{w}^{(1)} \in \mathbb{R}^n, b^{(1)} \in \mathbb{R}, \boldsymbol{\xi}^{(1)} \in \mathbb{R}_+^{N_B}$$

(TWSSVM-2)

$$Min \quad \frac{1}{2}\sum_{i=1}^{N_B}\left(\boldsymbol{w}^{(2)^T}\boldsymbol{x}_B^j + b^{(2)}\right)^2 + C_2 \sum_{i=1}^{N_A}\xi_i^{(2)}$$

$$s.t. \quad y_i\left(\boldsymbol{w}^{(2)^T}\boldsymbol{x}_A^i + b^{(2)}\right) \geq 1 - \xi_i^{(2)}, \; i = 1, \ldots, N_A$$

$$\boldsymbol{w}^{(2)} \in \mathbb{R}^n, b^{(2)} \in \mathbb{R}, \boldsymbol{\xi}^{(2)} \in \mathbb{R}_+^{N_A}$$

Observations

- Both (TWSSVM-1) and (TWSSVM-2) are convex quadratic programming problems.
- Compared to LSSVM, (TWSSVM-1) and (TWSSVM-2) have the same number of variables but fewer constraints.
- The total complexity of solving two smaller QPs is about ¼ of that solving LSSVM.
- The accuracy of TWSSVM is no inferior to that of LSSVM.

Example

- From reference [2]

Training Times (in Seconds)

Data Set	TWSVM (EXE file)	TWSVM (DLL file)	SVM (DLL file)
Hepatitis (155×19)	4.37	4.85	12.7
Sonar (208×60)	4.62	6.64	24.9
Heart-statlog (270×14)	4.72	11.3	50.9
Heart-c (303×14)	8.37	14.92	68.2
Ionosphere (351×34)	9.93	25.9	102.2
Votes (435×16)	12.8	45.8	189.4
Australian (690×14)	37.4	142.1	799.2
Pima-Indian (768×8)	56.9	231.5	1078.6
CMC (1473×9)	63.4	1737.9	6827.8

Data Set	TWSVM	GEPSVM	SVM
Heart-statlog (270×14)	84.44±4.32	84.81±3.87	84.07±4.40
Heart-c (303×14)	83.80±5.53	84.44±5.27	82.82±5.15
Hepatitis (155×19)	80.79±12.24	58.29±19.07	80.00±8.30
Ionosphere (351×34)	88.03±2.81	75.19±5.50	86.04 ±2.37
Sonar (208×60)	77.26±10.10	66.76±10.75	79.79±5.31
Votes (435×16)	96.08±3.29	91.93±3.18	94.50±2.71
Pima-Indian(768×8)	73.70±3.97	74.60±5.07	76.68±2.90
Australian (690×14)	85.80±5.05	85.65±4.60	85.51±4.58
CMC (1473×9)	67.28±2.21	65.99±2.30	67.82±2.63

Accuracies have been indicated as percentages.

TWSSVM

- TWSSVM classifier

$$class_{TWSSVM}(\boldsymbol{x}) = \text{arg}min\,\{|f_A(\boldsymbol{x})|, |f_B(\boldsymbol{x})|\}$$

- Primal version TWSSVM

$$f_A(\boldsymbol{x}) = {\boldsymbol{w}^{(1)}}^T \boldsymbol{x} + b^{(1)}$$

$$f_B(\boldsymbol{x}) = {\boldsymbol{w}^{(2)}}^T \boldsymbol{x} + b^{(2)}$$

Dual TWSSVM and kernel-based TWSSVM

- Basic approach:

1. Following the same procedure of finding dual LSSVM, we can derive dual TWSSVM.

2. Following the same procedure of finding kernel-based LSSVM, we can derive kernel-based TWSSVM.

TWSSVM model – vector form

(TWSSVM-1)

$$
\begin{aligned}
Min \quad & \frac{1}{2}\left\|X_A \boldsymbol{w}^{(1)} + \boldsymbol{e}_A b^{(1)}\right\|_2^2 + \mathcal{C}_1 \boldsymbol{e}_B^T \boldsymbol{\xi}^{(1)} \\
s.t. \quad & (-1)\left(X_B \boldsymbol{w}^{(1)} + \boldsymbol{e}_B \ b^{(1)}\right) \geq \boldsymbol{e}_B - \boldsymbol{\xi}^{(1)} \\
& \boldsymbol{w}^{(1)} \in \mathbb{R}^n, b^{(1)} \in \mathbb{R}, \boldsymbol{\xi}^{(1)} \in \mathbb{R}_+^{N_B}
\end{aligned}
$$

(TWSSVM-2)

$$
\begin{aligned}
Min \quad & \frac{1}{2}\left\|X_B \boldsymbol{w}^{(2)} + \boldsymbol{e}_B b^{(2)}\right\|_2^2 + \mathcal{C}_2 \boldsymbol{e}_A^T \boldsymbol{\xi}^{(2)} \\
s.t. \quad & (+1)\left(X_A \boldsymbol{w}^{(2)} + \boldsymbol{e}_B \ b^{(2)}\right) \geq \boldsymbol{e}_A - \boldsymbol{\xi}^{(2)} \\
& \boldsymbol{w}^{(2)} \in \mathbb{R}^n, b^{(2)} \in \mathbb{R}, \boldsymbol{\xi}^{(\mathbf{2})} \in \mathbb{R}_+^{N_A}
\end{aligned}
$$

Dual TWSSVM-1 model

- Lagrangian

$$L\left(\boldsymbol{w}^{(1)}, b^{(1)}, \boldsymbol{\xi}^{(1)}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right) = \frac{1}{2}\left\|X_A \boldsymbol{w}^{(1)} + \boldsymbol{e}_A b^{(1)}\right\|_2^2 + C_1 \boldsymbol{e}_B^T \boldsymbol{\xi}^{(1)} + \boldsymbol{\alpha}^T(\boldsymbol{e}_B - \boldsymbol{\xi}^{(1)} + X_B \boldsymbol{w}^{(1)} + \boldsymbol{e}_B\ b^{(1)}) - \boldsymbol{\theta}^T \boldsymbol{\xi}^{(1)}$$

- K-K-T conditions

(i) $X_A^T\left(X_A \boldsymbol{w}^{(1)} + \boldsymbol{e}_A b^{(1)}\right) + X_B^T \boldsymbol{\alpha} = 0;$

(ii) $\boldsymbol{e}_A^T(X_A \boldsymbol{w}^{(1)} + \boldsymbol{e}_A b^{(1)}) + \boldsymbol{e}_B^T \boldsymbol{\alpha} = 0;$

(iii) $C_1 \boldsymbol{e}_B - \boldsymbol{\alpha} - \boldsymbol{\theta} = 0;$

(iv) $-\left(X_B \boldsymbol{w}^{(1)} + \boldsymbol{e}_B\ b^{(1)}\right) + \boldsymbol{\xi}^{(1)} \geq \boldsymbol{e}_B\ ,\ \boldsymbol{\xi}^{(1)} \geq 0;$

(v) $\boldsymbol{\alpha}^T(-\left(X_B \boldsymbol{w}^{(1)} + \boldsymbol{e}_B\ b^{(1)}\right) + \boldsymbol{\xi}^{(1)} - \boldsymbol{e}_B\ ) = 0, \boldsymbol{\theta}^T\ \boldsymbol{\xi}^{(1)} = 0;$

(vi) $\boldsymbol{\alpha} \geq 0, \boldsymbol{\theta} \geq 0\ .$

Dual TWSSVM-1 model

(iii) says $\mathcal{C}_1 \boldsymbol{e}_B \geq \boldsymbol{\alpha} \geq 0,$

$$\Leftrightarrow 0 \leq \alpha_j \leq \mathcal{C}_1 \text{ for } j = 1, \dots, N_B$$

(i)+(ii) says

$$\begin{pmatrix} X_A^T \\ \boldsymbol{e}_A^T \end{pmatrix} (X_A, \boldsymbol{e}_A) \begin{pmatrix} \boldsymbol{w}^{(1)} \\ b^{(1)} \end{pmatrix} + \begin{pmatrix} X_B^T \\ \boldsymbol{e}_B^T \end{pmatrix} \boldsymbol{\alpha} = \mathbf{0}$$

$$\Leftrightarrow \quad \hat{X}_A^T \hat{X}_A \boldsymbol{u}^{(1)} + \hat{X}_B^T \ \boldsymbol{\alpha} = \mathbf{0}$$

$$\Leftrightarrow \quad \boldsymbol{u}^{(1)} = -\left(\hat{X}_A^T \hat{X}_A\right)^{-1} \hat{X}_B^T \boldsymbol{\alpha} \ \left[\text{practically, use } \left(\hat{X}_A^T \hat{X}_A + \varepsilon I\right)^{-1}\right]$$

- dual objective function

$$h(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \quad Min_{\boldsymbol{w}^{(1)} \in \mathbb{R}^n, b^{(1)} \in \mathbb{R}, \ \boldsymbol{\xi}^{(1)} \in \mathbb{R}_+^{N_B}} L\left(\boldsymbol{w}^{(1)}, b^{(1)}, \boldsymbol{\zeta}^{(1)}, \boldsymbol{\alpha}, \boldsymbol{\theta}\right)$$

$$= -\frac{1}{2} \boldsymbol{\alpha}^T \hat{X}_B \left(\hat{X}_A^T \hat{X}_A\right)^{-1} \hat{X}_B^T \ \boldsymbol{\alpha} + \boldsymbol{e}_B^T \boldsymbol{\alpha}$$

Dual TWSSVM model

(DTWSSVM-1)

$$Max \quad -\frac{1}{2}\boldsymbol{\alpha}^T \hat{X}_B(\hat{X}_A^T\hat{X}_A)^{-1}\hat{X}_B^T\boldsymbol{\alpha} + \boldsymbol{e_B^T\alpha}$$

$$s.t. \quad 0 \le \alpha_j \le \mathcal{C}_1 \text{ for } j = 1, \dots, N_B$$

dual-primal conversion: $\begin{pmatrix} \boldsymbol{w}^{(1)} \\ b^{(1)} \end{pmatrix} = \boldsymbol{u}^{(1)} = -(\hat{X}_A^T\hat{X}_A)^{-1}\hat{X}_B^T\boldsymbol{\alpha}$

*A simple convex quadratic programming problem.

Similarly, we have

(DTWSSVM-2)

$$Max \quad -\frac{1}{2}\boldsymbol{\gamma}^T \hat{X}_A(\hat{X}_B^T\hat{X}_B)^{-1}\hat{X}_A^T\boldsymbol{\gamma} + \boldsymbol{e_A^T\gamma}$$

$$s.t. \quad 0 \le \gamma_i \le \mathcal{C}_2 \text{ for } i = 1, \dots, N_A$$

dual-primal conversion: $\begin{pmatrix} \boldsymbol{w}^{(2)} \\ b^{(2)} \end{pmatrix} = \boldsymbol{u}^{(2)} = -(\hat{X}_B^T\hat{X}_B)^{-1}\hat{X}_A^T\boldsymbol{\gamma}$

DTWSSVM

- TWSSVM classifier

$$class_{TWSSVM}(\boldsymbol{x}) = \text{arg}min\,\{|f_A(\boldsymbol{x})|, |f_B(\boldsymbol{x})|\}$$

- Dual version TWSSVM

Denote $\hat{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix}$

$$f_A(\boldsymbol{x}) = -\left(\left(\hat{X}_A^T \hat{X}_A\right)^{-1} \hat{X}_B^T \boldsymbol{\alpha}\right)^T \hat{\boldsymbol{x}}$$

$$f_B(\boldsymbol{x}) = -\left(\left(\hat{X}_B^T \hat{X}_B\right)^{-1} \hat{X}_A^T \boldsymbol{\gamma}\right)^T \hat{\boldsymbol{x}}$$

Kernel-based twin soft SVM - KTWSSVM

- Using a *feature map* $\phi(\cdot) : \mathbb{R}^n \to \mathbb{R}^l \ (l \geq n)$ to transform the problem to a higher dimensional space for linear separability.

(KTWSSVM-1)

$$\text{Min} \quad \frac{1}{2}\sum_{i=1}^{N_A}\left(\boldsymbol{w}^{(1)^T}\phi(\boldsymbol{x}_A^i) + b^{(1)}\right)^2 + \ C_1 \ \sum_{j=1}^{N_B}\xi_j^{(1)}$$

$$\text{s.t.} \quad y_j\left(\boldsymbol{w}^{(1)^T}\phi(\boldsymbol{x}_B^j) + b^{(1)}\right) \geq 1 - \xi_j^{(1)}, \ j = 1, \dots, N_B$$

$$\boldsymbol{w}^{(1)} \in \mathbb{R}^l, b^{(1)} \in \mathbb{R}, \boldsymbol{\xi}^{(1)} \in \mathbb{R}_+^{N_B}$$

(KTWSSVM-2)

$$\text{Min} \quad \frac{1}{2}\sum_{i=1}^{N_B}\left(\boldsymbol{w}^{(2)^T}\phi(\boldsymbol{x}_B^j) + b^{(2)}\right)^2 + \ C_2 \ \sum_{i=1}^{N_A}\xi_i^{(2)}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^{(2)^T}\phi(\boldsymbol{x}_A^i) + b^{(2)}\right) \geq 1 - \xi_i^{(2)}, \ i = 1, \dots, N_A$$

$$\boldsymbol{w}^{(2)} \in \mathbb{R}^l, b^{(2)} \in \mathbb{R}, \boldsymbol{\xi}^{(\mathbf{2})} \in \mathbb{R}_+^{N_A}$$

Dual kernel-based TWSSVM model

(DKTWSSVM-1)

$$Max \ -\frac{1}{2}\boldsymbol{\alpha}^T\phi(X_B)(\phi(X_A)^{\mathrm{T}}\phi(X_A))^{-1}\phi(X_B)^T\boldsymbol{\alpha}+\boldsymbol{e}_{\boldsymbol{B}}^{\boldsymbol{T}}\boldsymbol{\alpha}$$

$$s.t. \quad 0 \le \alpha_j \le \ \mathcal{C}_1 \text{ for } j = 1, \dots, N_B$$

** Kernel matrix $K_A \triangleq \phi(X_A)^{\mathrm{T}}\phi(X_A)$

(DKTWSSVM-2)

$$Max \ -\frac{1}{2}\boldsymbol{\gamma}^T\phi(X_A)(\phi(X_B)^{\mathrm{T}}\phi(X_B))^{-1}\phi(X_A)^T\boldsymbol{\gamma}+\boldsymbol{e}_{\boldsymbol{A}}^{\boldsymbol{T}}\boldsymbol{\gamma}$$

$$s.t. \quad 0 \le \gamma_i \le \ \mathcal{C}_2 \text{ for } i = 1, \dots, N_A$$

** Kernel matrix $K_B \triangleq \phi(X_B)^{\mathrm{T}}\phi(X_B)$

*$\phi(X_A)^{\mathrm{T}} \in \boldsymbol{M}^{(l+1)\times N_A}$ formed by $\{\phi(x_A^i)\}$ with the last row being all 1's.

$\phi(X_B)^{\mathrm{T}} \in \boldsymbol{M}^{(l+1)\times N_B}$ form by $\{\phi\left(x_B^j\right)\}$ with the last row being all 1's.