Conditional probability estimation and sufficient dimension reduction with support matrix machine

Chanwoo Lee
Joint work with Miaoyan Wang

Department of Statistics
University of Wisconsin - Madison

May 28, 2020

Introduction: Support Vector Machine

- Given a set of training data $\{(\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\} : n = 1, \ldots, N\}$, we would like to learn a model with low error on the training data.
- One successful approach is a support vector machine (SVM).
- SVM finds an optimal hyperplane $\{\boldsymbol{x} : f(\boldsymbol{x}; \alpha, \boldsymbol{\beta}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = 0\}$ that separate the training data according to the labels.
- A classification rule induced by $f(\boldsymbol{x}; \alpha, \boldsymbol{\beta})$ is

$$g(\boldsymbol{x}; \alpha, \boldsymbol{\beta}) = \mathsf{sign}\left(f(\boldsymbol{x}; \alpha, \boldsymbol{\beta})\right) = \mathsf{sign}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

Introduction: SVM estimation

- The linear SVM solves

$$(\hat{\alpha}_N, \hat{\boldsymbol{\beta}}_N^T)^T = \operatorname*{arg\,min}_{\alpha, \boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2 + \frac{\lambda}{N} \sum_{n=1}^{N} \left|1 - y_i(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_n)\right|_+$$

- Using the duality, it can be shown that

$$\hat{\boldsymbol{\beta}}_N = \sum_{n=1}^{N} c_n \boldsymbol{x}_n \quad \text{where } c_n \in \mathbb{R}.$$

The case where predictor variables are matrices or higher order tensors

- In many classification problems, the input feature are naturally expressed as matrices or tensors rather than vectors.
  ex) electroencephalogram (EEG), image classification.
- SVM can not make use of the structure information of the original feature matrix.
- New method is needed, which can consider the correlation between columns and rows in the feature matrix.

Main goals

- From a given set of training data

$$\{(\boldsymbol{X}_n, y_m) \in \mathbb{R}^{d_1 \times d_2} \times \{-1, +1\} : n = 1, \ldots, N\},$$

we want to develop estimation methods for

1. Classifier (Support Matrix Machine): $\boldsymbol{g} : \mathbb{R}^{d_1 \times d_2} \to \{-1, +1\}$
2. Conditional probability: $\mathbb{P}(Y = 1 | \boldsymbol{X})$
3. Sufficient dimension reduction: $Y \perp\!\!\!\perp \boldsymbol{X} | T(\boldsymbol{X})$

- For 2 and 3, we will focus on linear estimation.

1. Support Matrix Machine (SMM): Linear model

- SMM finds an optimal hyperplane that separate the training data,

$$\{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2} : \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{B}, \alpha) = \alpha + \langle \boldsymbol{B}, \boldsymbol{X} \rangle\}, \tag{1}$$

  where $\langle \boldsymbol{B}, \boldsymbol{X} \rangle = \mathrm{tr}(\boldsymbol{B}^T \boldsymbol{X})$.

- A classification rule induced by $f(\boldsymbol{X}; \boldsymbol{B}, \alpha)$ is

$$\boldsymbol{g}(\boldsymbol{X}; \boldsymbol{B}, \alpha) = \mathrm{sign}\left(f(\boldsymbol{X}; \boldsymbol{B}, \alpha)\right) = \mathrm{sign}(\alpha + \langle \boldsymbol{B}, \boldsymbol{X} \rangle).$$

- When matrix $\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}$ is full rank, (1) is the same as SVM.
- To exploit the correlation information of predictor $\boldsymbol{X}$, we impose low rank structure on $\boldsymbol{B}$ as

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T \quad \text{where } \boldsymbol{U} \in \mathbb{R}^{d_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{d_1 \times r}$$

1. SMM estimation: Linear model

- The linear SMM solves

$$(\hat{\alpha}_N, \hat{\boldsymbol{U}}_N, \hat{\boldsymbol{V}}_N) = \operatorname*{arg\,min}_{\alpha, \boldsymbol{U}, \boldsymbol{V}} \|\boldsymbol{U}\boldsymbol{V}^T\|^2 + \frac{\lambda}{N}\sum_{n=1}^{N} \left|1 - y_n(\alpha + \langle \boldsymbol{U}\boldsymbol{V}^T, \boldsymbol{X}\rangle)\right|_+ . \tag{2}$$

- We can optimize (2) with a coordinate descent algorithm updating $\boldsymbol{U}$ holding $\boldsymbol{V}$ fixed and vice versa.
- Using the duality, it can be shown that

$$\hat{\boldsymbol{B}}_N = \hat{\boldsymbol{U}}_N \hat{\boldsymbol{V}}_N^T = \sum_{n=1}^{N} c_n H_{\hat{\boldsymbol{U}}_N} \boldsymbol{X}_n H_{\hat{\boldsymbol{V}}_N} \quad \text{where } H_A = A(A^T A)^{-1} A^T \tag{3}$$

- (3) gives us intuition how SMM uses information about the correlation among columns or rows.

1. SMM: Nonlinear model

- Linear boundaries in the enlarged space can translate to nonlinear boundaries in the original space.
- We map original space to enlarged space with feature mapping

$$\boldsymbol{h} : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}^{d_1' \times d_2}.$$

- Nonlinear SMM finds an optimal hyperplane in enlarged space

$$\{\boldsymbol{h}(\boldsymbol{X}) \in \mathbb{R}^{d_1' \times d_2} : \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{U}, \boldsymbol{V}, \alpha) = \alpha + \langle \boldsymbol{U}\boldsymbol{V}^T, \boldsymbol{h}(\boldsymbol{X}) \rangle\}.$$

1. SMM: Nonlinear model

- It can be shown that the solution function $\boldsymbol{f}(\boldsymbol{X})$ can be written as

$$
\begin{aligned}
\boldsymbol{f}(\boldsymbol{X};\boldsymbol{U},\boldsymbol{V},\alpha) &= \alpha + \langle \boldsymbol{U}\boldsymbol{V}^T, \boldsymbol{h}(\boldsymbol{X})\rangle \\
&= \alpha + \sum_{i=1}^{N} c_i \operatorname{tr}\left(H_{\boldsymbol{V}} \boldsymbol{h}(\boldsymbol{X})^T \boldsymbol{h}(\boldsymbol{X}_i)\right) \\
&= \alpha + \sum_{i=1}^{N} c_i \operatorname{tr}\left(H_{\boldsymbol{V}} \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_i)\right),
\end{aligned}
$$

  where we define $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}') = \boldsymbol{h}(\boldsymbol{X})^T \boldsymbol{h}(\boldsymbol{X}')$ and $c_i \in \mathbb{R}$.
- In fact, we need not specify $\boldsymbol{h}(\boldsymbol{X})$ at all, but require only knowledge of $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}')$.

1. SMM: Nonlinear kernel functions

- There are some kernels that might be used often,

$$\begin{aligned} \text{Linear:} \quad & \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}') = \boldsymbol{X}^T \boldsymbol{X}', \\ \text{Polynomial:} \quad & \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}') = (\boldsymbol{X}^T \boldsymbol{X}' + \boldsymbol{I}_n)^d, \\ \text{Radial:} \quad & \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}') = \exp\left((\boldsymbol{X} - \boldsymbol{X}')^T (\boldsymbol{X} - \boldsymbol{X}')/\sigma\right). \end{aligned}$$

- We transform $\boldsymbol{X}_i^* = \begin{pmatrix} 0 & \boldsymbol{X}_i^T \\ \boldsymbol{X}_i & 0 \end{pmatrix}$ for symmetric adjustment.

2. Conditional probability estimation

- We estimate conditional probability $\mathbb{P}(Y=1|\boldsymbol{X})$ based on SMM inference where $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$.
- SMM classifier can be fit in the following regularization frame work with $\mathcal{F} = \{\boldsymbol{f}(\boldsymbol{X};\boldsymbol{B},\alpha) = \alpha + \langle \boldsymbol{B}, \boldsymbol{X} \rangle : \alpha \in \mathbb{R}, \boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}\}$ and $J(\boldsymbol{f}(\boldsymbol{X};\boldsymbol{B},\alpha)) = \|\boldsymbol{B}\|^2$.

  inner function;

  sign(f(x)) = G(x) (classifier; classification rule)

  f: argument function?

$$\min_{\boldsymbol{f} \in \mathcal{F}} J(\boldsymbol{f}) + \frac{\lambda}{N} \sum_{n=1}^{N} \omega_\pi(Y_n) \left|1 - Y_n \boldsymbol{f}(\boldsymbol{X}_n)\right|_+ , \tag{4}$$

  where $\omega_\pi(Y) = 1 - \pi$ if $Y = 1$ and π if $Y = -1$ with a weight $\pi \in (0,1)$.
- We base our estimation method on the following theorem.

Theorem 1

When $N \to \infty$, minimizing (4) with respect to $\boldsymbol{f}$ targets directly at sign $[\mathbb{P}(Y=1|\boldsymbol{X}) - \pi]$

as a function of X: eta(X);

2. Conditional probability estimation: Algorithm

- From a set of training data $\{(\boldsymbol{X}_n, Y_n)\}_{n=1}^{N}$, we estimate $\mathbb{P}(Y = 1|\boldsymbol{X})$ for new predictor $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ as follows.
  1. Initialize $\pi_h = (h-1)/H$, for $h = 1, \ldots, H+1$.
  2. Train a weighted margin classifgier for π_h as in (4), for $h = 1, \ldots H+1$.
  3. Estimate labels of $\boldsymbol{X}$ by sign $\left(\hat{\boldsymbol{f}}_{\pi_h}(\boldsymbol{X})\right)$.
  4. Sort sign $\left(\hat{\boldsymbol{f}}_{\pi_h}(\boldsymbol{X})\right)$, $h = 1, \ldots, H+1$, and obtain estimated probability $\hat{\mathbb{P}}(Y = 1|\boldsymbol{X})$ as

$$\frac{1}{2}\left(\arg\max_{\pi_h}\{\text{sign}(\hat{\boldsymbol{f}}_{\pi_h}(\boldsymbol{X})) = 1\} + \arg\max_{\pi_h}\{\text{sign}(\hat{\boldsymbol{f}}_{\pi_h}(\boldsymbol{X})) = -1\}\right).$$

3. Sufficient Dimension Reduction

- For a matrix predictor $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, sufficient dimension reduction assumes that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V}, \tag{5}$$

  where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times k_1}, \boldsymbol{V} \in \mathbb{R}^{d_2 \times k_2}$.

- We can equivalently express (5) as

$$Y \perp\!\!\!\perp \boldsymbol{X} | \left\{ \left\langle \boldsymbol{u}_i \boldsymbol{v}_j^T, \boldsymbol{X} \right\rangle \right\}_{i \in [k_1], j \in [k_2]}$$

  where $\boldsymbol{u}_i$ is i-th column of $\boldsymbol{U}$ and $\boldsymbol{v}_j$ is j-th column of $\boldsymbol{V}$.

- The central subspace in matrix case is defined as

$$S_{Y|\boldsymbol{X}} = \bigcap_{\{(\boldsymbol{U},\boldsymbol{V}): Y \perp\!\!\!\perp X | \boldsymbol{X} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V}\}} \text{span}(\boldsymbol{U}) \times \text{span}(\boldsymbol{V}),$$

3. Sufficient Dimension Reduction

- We can consider the linear principal weighted support matrix machine

$$\Lambda_\pi(\boldsymbol{u}, \boldsymbol{v}) = \mathsf{Var}(\langle \boldsymbol{u}\boldsymbol{v}^T, \boldsymbol{X} \rangle) + \lambda \mathbb{E}\left\{ \omega_\pi(Y) \left| 1 - Y \boldsymbol{f}(\boldsymbol{X}; \boldsymbol{u}, \boldsymbol{v}, \alpha) \right|_+ \right\}, \quad (6)$$

where $\boldsymbol{f}(\boldsymbol{X}; \boldsymbol{u}, \boldsymbol{v}, \alpha) = \alpha + \langle \boldsymbol{u}\boldsymbol{v}^T, \boldsymbol{X} - \mathbb{E}(\boldsymbol{X}) \rangle$.

- Weighted SMM is a splecial case of (6).
(when $\mathbb{E}(\mathrm{Vec}(\boldsymbol{X})) = 0 \in \mathbb{R}^{d_1 d_2}, \quad \mathsf{cov}(\mathrm{Vec}\,(\boldsymbol{X})) = \boldsymbol{I}_{d_1 d_2}$)
- We base our estimation method on the following theorem.

Theorem 2 (Not verified yet)

Assume that $\mathbb{E}(\boldsymbol{X} | \boldsymbol{X} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V})$ is a linear function of $\boldsymbol{X} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V}$. Then for any given weight $\pi \in (0, 1)$, the optimizer $(\boldsymbol{u}_{0,\pi}, \boldsymbol{v}_{0,\pi})$ of (6) belongs to $S_{Y|\boldsymbol{X}}$ under (5).

3. Sufficient Dimension Reduction: Algorithm

- The sampled version of Λ_π in (6) is,

$$\hat{\Lambda}_{N,\pi} = \mathrm{Vec}\left(\boldsymbol{u}\boldsymbol{v}^T\right)^T \hat{\boldsymbol{\Sigma}}_{\mathbf{N}} \mathrm{Vec}\left(\boldsymbol{u}\boldsymbol{v}^T\right) + \frac{\lambda}{N}\sum_{i=1}^{N} \omega_\pi(Y_i)\left(1 - Y_i \hat{f}_N(\boldsymbol{X}_i; \boldsymbol{u}, \boldsymbol{v}, \alpha)\right)_+, \tag{7}$$

- From standardization for $\{\mathrm{Vec}(\boldsymbol{X}_n)\}_{n=1}^N$ and reparameterization, (7) is expressed as regular weighted SMM objective function.
- We obtain the optimizer $(\hat{\boldsymbol{u}}_{N,\pi}, \hat{\boldsymbol{v}}_{N,\pi})$ with the same algorithm in Section 2.

3. Sufficient Dimension Reduction: Algorithm

- From a set of training data $\{(\boldsymbol{X}_n, Y_n)\}_{n=1}^{N}$, we estimate the central subspace $S_{Y|\boldsymbol{X}}$ as follows
  1. Initialize $\pi_h = (h-1)/H$, for $h = 1, \ldots, H+1$.
  2. Given a grid $0 < \pi_1 < \cdots < \pi_H < 1$, we obtained H-candidates $\{\hat{\boldsymbol{u}}_{n,\pi_h} \hat{\boldsymbol{v}}_{n,\pi_h}^T\}_{h=1}^{H}$ for the central subspace.
  3. Obtain the candidate tensor $\hat{\mathcal{M}} \in \mathbb{R}^{H \times d_1 \times d_2}$ such that $\hat{\mathcal{M}}_{h\cdot\cdot} = \hat{\boldsymbol{u}}_{n,\pi_h} \hat{\boldsymbol{v}}_{n,\pi_h}^T$
  4. Obtain order-3 SVD as,

$$\hat{\mathcal{M}} = \hat{\mathcal{C}} \times_1 \hat{\boldsymbol{U}}_1 \times_2 \hat{\boldsymbol{U}}_2 \times_3 \hat{\boldsymbol{U}}_3,$$

     where $\mathcal{C} \in \mathbb{R}^{H \times d_1 \times d_2}$,$\boldsymbol{U}_1 \in \mathbb{R}^{H \times H}$,$\boldsymbol{U}_2 \in \mathbb{R}^{d_2 \times d_2}$, and $\boldsymbol{U}_3 \in \mathbb{R}^{d_3 \times d_3}$.
  5. Estimate the central subspace as

$$\hat{S}_{Y|\boldsymbol{X}} = \{[\hat{\boldsymbol{U}}_2]_i\}_{i=1}^{k_1} \times \{[\hat{\boldsymbol{U}}_3]_i\}_{i=1}^{k_2}.$$

We can reduce the dimension of $\boldsymbol{X}$ as $\{\boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v}$ where $(\boldsymbol{u}, \boldsymbol{v}) \in \hat{S}_{Y|\boldsymbol{X}}\}$.