A Demo of Reduced Gradient Descent Optimization

Jiyuan Liu

June 13, 2020

Date Performed:	June 13, 2020
Partners:	None
Supervisor:	Xinwang Liu

1 Objective

The primal problem is

$$\begin{aligned}
&\min_{R,\mathbf{c}',\boldsymbol{\xi}} R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\
&s.t. \left\|\phi\left(\mathbf{x}_i\right) - \mathbf{c}'\right\|^2 \leq R^2 + \xi_i,\ \xi_i \geq 0
\end{aligned} \tag{1}$$

Defining

$$\Phi_1(\cdot) = \begin{bmatrix} \phi_1(\cdot) \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \Phi_2(\cdot) = \begin{bmatrix} 0 \\ \phi_2(\cdot) \\ \vdots \\ 0 \end{bmatrix}, \ldots, \Phi_m(\cdot) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \phi_m(\cdot) \end{bmatrix}, \tag{2}$$

and assuming the optimal kernel mapping $\Psi(\cdot) = \sum_{p=1}^{m} \gamma_p \Phi_p(\cdot)$ s.t. $\sum_{i=1}^{m} \gamma = 1$, the multiple kernel support vector data description is able to be fomulated as

$$\begin{aligned}
&\min_{R,\mathbf{c},\boldsymbol{\xi},\boldsymbol{\gamma}} R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\
&s.t. \left\|\Psi\left(\mathbf{x}_i\right) - \mathbf{c}\right\|^2 \leq R^2 + \xi_i,\ \xi_i \geq 0,\ \sum_{p=1}^{m}\gamma_p = 1,\ \gamma_p \geq 0
\end{aligned} \tag{3}$$

Two main differences between Eq. (1) and (3) are:

- Compared with $\phi(\mathbf{x})$ in Eq. (1), $\Psi(\mathbf{x})$ in Eq. (3) is a linear combination of pre-specified mappings;
- the dimension of $\mathbf{c}$ in Eq. (3) is $\sum_{p=1} dim\left(\Phi(x_p)\right)$ which differs from $\mathbf{c}'$ in Eq. (1).

The dual problem of Eq. (3) can be represented as

$$\begin{aligned}
\mathcal{L} =& R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[R^2 + \xi_i - \|\Psi(\mathbf{x}_i) - \mathbf{c}\|^2\right] - \sum_{i=1}^{n}\beta_i\xi_i \\
&+ \theta\left(\sum_{p=1}^{m}\gamma_p - 1\right) - \sum_{p=0}^{m}\mu_p\gamma_p
\end{aligned} \tag{4}$$

Setting partial derivatives to zero as

$$
\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial R} = 0: \ \sum_{i=1}^{n} \alpha_i = 0 \\
&\frac{\partial \mathcal{L}}{\partial \mathbf{c}} = 0: \ \mathbf{c} = \sum_{i=1}^{n} \alpha_i \Psi(\mathbf{x}_i) \\
&\frac{\partial \mathcal{L}}{\partial \xi_i} = 0: \ \frac{1}{\nu n} - \alpha_i - \xi_i = 0 \\
&\frac{\partial \mathcal{L}}{\partial \gamma_p} = 0:
\end{aligned}
\tag{5}
$$

we can get

$$
L(\alpha_i, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \alpha_i \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j). \tag{6}
$$

in which $\Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j) = \sum_{p=1}^{m} \gamma_p^2 \Phi_p(\mathbf{x}_i) \cdot \Phi_p(\mathbf{x}_j) = \sum_{p=1}^{m} \gamma_p^2 \phi_p(\mathbf{x}_i) \cdot \phi_p(\mathbf{x}_j)$. With defining $\mathbf{K}_p$ as the kernel matrix specified by mapping $\phi_p(\cdot)$ and $\mathbf{K}_{\boldsymbol{\gamma}} = \sum_{p=1}^{m} \gamma_p \mathbf{K}_p$, Eq. (6) can be termed as

$$
\begin{aligned}
&\min_{\boldsymbol{\gamma}} \max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}^\top \mathbf{K}_{\boldsymbol{\gamma}} \boldsymbol{\alpha} - \mathbf{f}^T \boldsymbol{\alpha} \\
&\qquad s.t. \ \mathbf{f} = diag(\mathbf{K}_{\boldsymbol{\gamma}}), \ \sum_{p=1}^{m} \gamma_p^2 = 1, \ 0 \leq \gamma_p \leq 1, \\
&\qquad \sum_{i=1}^{n} \alpha_i = 1, \ 0 \leq \alpha_i \leq \frac{1}{\nu n}
\end{aligned}
\tag{7}
$$

2 Optimization

Opimizing c and R**.**
Fixing $\boldsymbol{\gamma}$, Eq. (12) is a Quadratic-Programming problem respected to $\boldsymbol{\alpha}$.

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}^\top \mathbf{K}_{\boldsymbol{\gamma}} \boldsymbol{\alpha} - \mathbf{f}^T \boldsymbol{\alpha} \\
&s.t. \ \mathbf{f} = diag(\mathbf{K}_{\boldsymbol{\gamma}}), \ \sum_{i=1}^{n} \alpha_i = 1, \ 0 \leq \alpha_i \leq \frac{1}{\nu n}
\end{aligned}
\tag{8}
$$

With $\boldsymbol{\alpha}$, the sphere center can be represented as

$$
\mathbf{c} = \sum_{i=1}^{n} \alpha_i \Psi(\mathbf{x}_i) \tag{9}
$$

R is the ν-quantile of distances from data points to center $\mathbf{c}$.

Optimizing $\boldsymbol{\gamma}$.

$$\begin{aligned} &\max_{\boldsymbol{\gamma}} \boldsymbol{\gamma}^\top \mathbf{V} \\ &s.t.\ V_p = diag(\mathbf{K}_p)^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \mathbf{K}_p \boldsymbol{\alpha},\ \sum_{p=1}^{m} \gamma_p^2 = 1 \end{aligned} \tag{10}$$

in which $\boldsymbol{\gamma}$ is solved as

$$\gamma_p = \frac{V_p}{\sqrt{V_1^2 + V_2^2 + \cdots + V_m^2}} \tag{11}$$

3 Prediction

A data point $\mathbf{z}$ is accepted as abnormality by:

$$\|\mathbf{z} - \mathbf{c}\|^2 = \Psi(\mathbf{z}) \cdot \Psi(\mathbf{z}) - 2\sum_{i=1}^{n} \alpha_i \Psi(\mathbf{z}) \cdot \Psi(\mathbf{x}_i) + \sum_{i,j=1}^{n} \alpha_i \alpha_j \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j) \geq R^2 \tag{12}$$

Tips: the 3-rd term can be calculated in advance and stored for further usage.

4 Kernel function

Linear:

$$k(x, y) = x^\top y + c \tag{13}$$

Polynomial:

$$k(x, y) = (ax^\top y + c)^d \tag{14}$$

Rbf:

$$k(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2}) \tag{15}$$

Laplacian:

$$k(x, y) = \exp(-\frac{\|x - y\|}{\sigma}) \tag{16}$$

Sigmoid:

$$k(x, y) = \tanh(ax^\top y + c) \tag{17}$$

Inverse multiquadric:

$$k(x, y) = \frac{1}{\sqrt{\|x - y\|^2 + c^2}} \tag{18}$$

Log:

$$k(x, y) = -\log(\|x - y\|^d + 1) \tag{19}$$

Cauthy:

$$k(x, y) = \frac{1}{\frac{\|x - y\|^2}{\sigma} + 1} \tag{20}$$

References