IDLSSVM-CIL Optimal w and b Values Calculation

Author: Atharva Mohite

In the study on Density-weighted Support Vector Machines for binary Class Imbalance Learning (DSVM-CIL)[1], the density weight notion is calculated as:

$$d_i = \rho(x_i) = 1 - \frac{d\left(x_i, x_i^k\right)}{max_{j\in train\_set}\, d\left(x_j, x_j^k\right)} \tag{1}$$

here x_i and x_i^k represents a datapoint and its k-NN ($k = 5$) datapoint, such that $0 < d_i < 1$

The diagonal matrix of (1) may be expressed as:

$$D = diag(d_i)$$

Note that the density weights are estimated in real space.

As proposed by Hazarika et al., Improved Density-weighted Least Squares SVM for binary Class Imbalance Learning (IDLSSVM-CIL) optimization problem is:

$$min\ \frac{1}{2}w^t w + \frac{C}{2}\sum_{i=1}^{m} d_i^2 \varphi_i^2$$

subject to $y_i\left(w^t \phi(x_i) + b\right) = 1 - \varphi_i\ ,\ i = 1, 2 \ldots m$

where w is the normal direction of the plane, b is a form of threshold for the hyperplane, $\phi(x)$ function maps the training data $x \in R^n$ to a higher dimension in feature space, φ_i is the i^{th} slack variable, $y_i \in \{-1, +1\}$ is the class label of i^{th} training sample, $C > 0$ is a user defined penalty parameter and m is the number of training samples.

The Langrage function $L(w, b, e, \alpha)$ is given as:

$$L(w, b, \varphi_i, \alpha_i) = \frac{1}{2}w^t w\ +\ \frac{C}{2}\sum_{i=1}^{m} d_i^2 \varphi_i^2 - \sum_{i=1}^{m} \alpha_i \left\{ y_i\left(w^t \phi(x_i) + b\right) - 1 + \varphi_i + \varepsilon \right\} \tag{2}$$

where α_i are Lagrange multipliers, which can be either positive or negative due to the equality constraints as follows from the Karush–Kuhn–Tucker (KKT) conditions.[2]

The conditions for optimality are obtained by equating the derivatives of the Langrage function (2) wrt w, b, φ_i and α_i to zero.

$$\frac{\partial L}{\partial b} = 0\ \rightarrow w = \sum_{i=1}^{m} \alpha_i y_i \phi(x_i) \tag{3}$$

$$\frac{\partial L}{\partial w} = 0\ \rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{4}$$

$$\frac{\partial L}{\partial \varphi_i} = 0\ \rightarrow \varphi_i = \frac{\alpha_i}{C d_i^2}\ ,\ i = 1, 2 \ldots m \tag{5}$$

$$\frac{\partial L}{\partial \alpha_i} = 0\ \rightarrow y_i\left(w^t \phi(x_i) + b\right) + \varphi_i = 1\ ,\ i = 1, 2 \ldots m \tag{6}$$

Substituting w and φ_i from (3) and (5) in (6):

$$y_i\left(\alpha_i \sum_{j=1}^{m} y_j \phi(x_i)^t \phi(x_j) + b\right) + \frac{\alpha_i}{C d_i^2} = 1\ ,\ i = 1, 2 \ldots m \tag{7}$$

Above one linear equation (4) and m linear equations contained in (7) can be written as:

$$\begin{bmatrix} 0 & y_1 & \cdots & y_m \\ y_1 & y_1^2 K(x_1, x_1) + \dfrac{1}{C d_1^2} + \varepsilon & \cdots & y_1 y_m K(x_1, x_m) + \dfrac{1}{C d_1^2} + \varepsilon \\ y_2 & y_2 y_1 K(x_2, x_1) + \dfrac{1}{C d_2^2} + \varepsilon & \cdots & y_2 y_m K(x_2, x_m) + \dfrac{1}{C d_2^2} + \varepsilon \\ \vdots & \vdots & \ddots & \vdots \\ y_m & y_m y_1 K(x_m, x_1) + \dfrac{1}{C d_m^2} + \varepsilon & \cdots & y_m^2 K(x_m, x_m) + \dfrac{1}{C d_m^2} + \varepsilon \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{8}$$

here $K(x_i, x_j) = \phi(x_i)^t \phi(x_j)$ is a non-linear kernel function (Radial Basis Function (RBF) in our case) and $\varepsilon > 0$ is the regularization term. Above matrix equation (8) can be written as:

$$A\vec{S} = B$$
$$\vec{S} = BA^{-1}$$

The first and the remaining m elements of vector $\vec{S}$ are the optimum values of b and the langrage multipliers α_i's.

The discriminative function $f(x)$, after the kernel trick, is:

$$f(x) = sign\left(\sum_{j=1}^{m} \alpha_j y_j K(x_j, x) + b\right)$$

---

[1] Hazarika, Barenya Bikash, and Deepak Gupta. "Density-weighted support vector machines for binary class imbalance learning." Neural Computing and Applications 33.9 (2021): 4243-4261.

[2] Fletcher, Roger. Practical methods of optimization. John Wiley & Sons, 2000.