Kernel SVM with RBF and k-NN Classification in Loan Risk Assessment

Tran Quoc Thai
Student ID: 2370759

April 10, 2025

1. Kernel SVM with Radial Basis Function (RBF)

Given a training dataset with feature vectors $\mathbf{x}_i = (\text{Age}, \text{CreditScore})$, the Radial Basis Function (RBF) kernel is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \tag{1}$$

where γ is the kernel parameter that controls the spread of the Gaussian function.

For the first three training samples and given $\gamma = 0.1$, we compute the kernel matrix:

$$K = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & K(\mathbf{x}_1, \mathbf{x}_3) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & K(\mathbf{x}_2, \mathbf{x}_3) \\ K(\mathbf{x}_3, \mathbf{x}_1) & K(\mathbf{x}_3, \mathbf{x}_2) & K(\mathbf{x}_3, \mathbf{x}_3) \end{bmatrix}. \tag{2}$$

Computing each entry using the Euclidean distance and applying the RBF kernel function, we obtain:

$$K = \begin{bmatrix} 1.000 & 0.778 & 0.856 \\ 0.778 & 1.000 & 0.692 \\ 0.856 & 0.692 & 1.000 \end{bmatrix}. \tag{3}$$

This transformation allows the SVM to classify non-linearly separable data by mapping them into a higher-dimensional space where a linear separator exists.

2. k-Nearest Neighbors Classification

To classify test sample T_1 with normalized values $\mathbf{x}_{T1} = (0.375, 0.583)$ using $k = 3$, we first compute Euclidean distances:

$$d(\mathbf{x}_{T1}, \mathbf{x}_i) = \sqrt{(x_{T1} - x_i)^2 + (y_{T1} - y_i)^2}. \tag{4}$$

Using the normalized training data:

$$d(\mathbf{x}_{T1}, \mathbf{x}_1) = \sqrt{(0.375 - 0.438)^2 + (0.583 - 0.750)^2} = 0.175, \tag{5}$$

$$d(\mathbf{x}_{T1}, \mathbf{x}_2) = \sqrt{(0.375 - 0.290)^2 + (0.583 - 0.417)^2} = 0.194, \tag{6}$$

$$d(\mathbf{x}_{T1}, \mathbf{x}_3) = \sqrt{(0.375 - 0.500)^2 + (0.583 - 0.833)^2} = 0.273. \tag{7}$$

Sorting the distances, the three nearest neighbors correspond to $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. If the majority class among these is "Low Risk," then T_1 is classified as "Low Risk."

The choice of k affects the decision boundary: smaller k values lead to more sensitive and complex boundaries, while larger k values provide smoother and more generalized decision regions.