Comparison of Distance-Weighted k-NN and Standard k-NN for Loan Risk Classification

Tran Quoc Thai
Student ID: 2370759

April 10, 2025

1 Introduction

This analysis compares the standard k-nearest neighbors (k-NN) and distance-weighted k-NN classification methods by evaluating the risk level of test sample T_1. We consider $k = 3$ and compute classification probabilities using both approaches.

2 Standard k-NN Classification

Given a test sample T_1 with normalized features:

$$T_1 = (0.375, 0.583)$$

We previously calculated Euclidean distances between T_1 and all training samples and identified the three nearest neighbors. Let N_1, N_2, N_3 be these nearest neighbors, and their respective risk classifications are:

$$N_1 = \text{High Risk},$$
$$N_2 = \text{Low Risk},$$
$$N_3 = \text{High Risk}.$$

Using majority voting, we classify T_1 as High Risk, since two of the three neighbors belong to the High Risk category.

3 Distance-Weighted k-NN Classification

In this approach, instead of simple majority voting, we weight each neighbor's contribution based on the inverse of its Euclidean distance to T_1. Let the dis-

tances be:

$$d_1 = 0.12,$$
$$d_2 = 0.15,$$
$$d_3 = 0.18.$$

The weights are computed as:

$$w_i = \frac{1}{d_i}, \quad i = 1, 2, 3.$$

Thus:

$$w_1 = \frac{1}{0.12} = 8.33,$$
$$w_2 = \frac{1}{0.15} = 6.67,$$
$$w_3 = \frac{1}{0.18} = 5.56.$$

Now, we calculate the probability of T_1 being High Risk:

$$P(\text{High Risk}) = \frac{w_1 + w_3}{w_1 + w_2 + w_3} = \frac{8.33 + 5.56}{8.33 + 6.67 + 5.56} = \frac{13.89}{20.56} \approx 0.676.$$

Similarly, the probability of T_1 being Low Risk is:

$$P(\text{Low Risk}) = \frac{w_2}{w_1 + w_2 + w_3} = \frac{6.67}{20.56} \approx 0.324.$$

Since $P(\text{High Risk}) > P(\text{Low Risk})$, we classify T_1 as High Risk.

4 Conclusion

The distance-weighted k-NN technique predicts high risk for T1 while refining the probability estimate making it less affected by distant neighbors. The approach proves more solid particularly for regions where features experience important fluctuations or when measurement errors lead to decreased accuracy.