4. Experiments:

Experiment 1:

Cosine distance is better. Because the data matrix is sparse, cosine distance is more suitable than Euclidean distance.

Experiment 2:

Whether for the training set or the test set, their accuracy will decrease as the value of k in KNN increases. When K is large enough, for example, the value of K is larger than the training sample size, then the model is for all samples. the prediction results are the same, and the model loses its predictive significance.

Experiment 3:

The second classifier is the worst, because the distribution of the classification labels is unbalanced, and the samples of Class 1 are much smaller than the samples of other classes.

5. Result Analysis:

Analysis 1:

The sample data is randomly sampled, 80 samples are selected as the training set, and the remaining samples are used as the test set. Then a KNN (1) model is established to predict the test set and calculate the error rate of the test set.

Repeat the above operation 20 times, you can get 20 test error rates, and use these statistics to calculate the confidence interval.

Analysis 2:

The method of calculating the confidence interval of the test error rate of KNN (45) is the same as the previous method of calculating the confidence interval of the test error rate of KNN (1), but the parameters of the model are different.

Comparing the test error rate confidence intervals of the two models, KNN (45) has a higher test error rate.

6. Hyperparameter Selection:

We choose random subsampling. First, shuffle the data randomly, select the first 600 samples as the training set, and the remaining samples as the test set.

We try the KNN model under different k values, and the k value is traversed from 1 to 50. For each KNN, make predictions on the test set and record the model training accuracy.

In this way, the KNN with the highest accuracy on the test set can be found and used as the best prediction model.