Medical Data Binary Classification

Introduction：
The Vertebral Column Data Set is a biomedical data set consisting of six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine. The aim of this exercise is to perform binary classification using the k-nearest neighbors' algorithm and to explore variants of KNN using different distance metrics.

Data Pre-Processing and Exploratory Data Analysis:

The independent variables in the dataset are represented by scatterplots, and boxplots are used to visualize the distribution of each independent variable. The first 70 rows of Class 0 and the first 140 rows of Class 1 are selected as the training set, and the rest of the data is used as the test set.

Classification using KNN on Vertebral Column Data Set:

The k-nearest neighbors algorithm with Euclidean metric is used for classification. The test data is tested using k nearest neighbors, and decisions are made by majority polling. The train and test errors are plotted in terms of k for $k \in \{208, 205, \ldots, 7, 4, 1, \}$ (in reverse order), and the optimal k is determined. The confusion matrix, true positive rate, true negative rate, precision, and F1-score are calculated when $k = k*$. The best test error rate is plotted against the size of the training set.

Exploring Variants of KNN:

The Euclidean metric is replaced with Minkowski Distance, Manhattan Distance, Chebyshev Distance, and Mahalanobis Distance. The test errors are summarized in a table, and the best k is selected for each distance metric. Weighted voting is used with Euclidean, Manhattan, and Chebyshev distances to determine the best test errors when $k \in \{1,6,11,16,...,196\}$.

Lowest Training Error Rate:

The lowest training error rate achieved in this homework is 0.08 when using mahalanobis distance.

Conclusion：

In this exercise, the Vertebral Column Data Set is used for binary classification using the k-nearest neighbors algorithm. Variants of KNN are explored using different distance metrics, and the lowest training error rate is reported.

Lingduo Luo | LINGDUOLUO109@GMAIL.COM