CS 532 Project Proposal

Name: Haotian Shi

Project Datasets

The data processed in this project is MNIST dataset, which is a widely used dataset for classification algorithm from National Institute of Standards and Technology (NIST). The source link is http://yann.lecun.com/exdb/mnist/. The training set consists of handwritten numbers from 250 different people, of which 50% are high school students and 50% are from the Census Bureau staff. The test set is also handwritten digital data in the same proportion. The training data set contains 60,000 samples, and the test data set contains 10,000 samples. Each picture in the MNIST data set consists of 28 x 28 pixels, and each pixel is represented by a gray value. The 28 x 28 pixels are expanded into a one-dimensional row vector with 784 values. These rows are the feature rows in the first array of a image. The second array (labels) contains the corresponding target variable, which is the class label of the handwritten number (integer 0-9).

Thus, the dimension of the training data 'X' is 60000 x 784, where each sample corresponds to a label from numbers 0-9. Similarly, the dimension for the test data 'X' is 10000 x 784. The classification problem is defined as: which number from 0 to 9 is most likely to be given a feature vector with 784 features.

Investigated Algorithms

The following three algorithms will be applied on the dataset, and corresponding parameters will be analyzed.

1. k-nearest neighbor

   parameters for experiment: number of neighbors K; weights (all points in each neighborhood are weighted equally or weight points by the inverse of their distance); algorithm (which algorithm is used to search for the nearest k points when building the KNN model), etc.

2. neural networks

   parameters for experiment: number of hidden layers, learning rate, momentum, activation function; optimization algorithm, regularization, etc.

3. SVM

   parameters for experiment: kernel, degree, gamma, decision function shape, etc.

Validation task

The dataset will be divided into multiple subsets for cross-validation, and the validation results are averaged over the rounds to better estimate the model's predictive performance.

Confusion matrix, precision and recall will be used to evaluate the model performance. The performance of each algorithm will be compared and the basic principle behind the algorithm will be analyzed to discuss the evaluation results.

Schedule

Week	Tasks
10/19-10/25	Understanding the basic principles of the three algorithms
10/26-11/01	Dataset processing, developing training environment

11/02-11/08	Learning algorithm investigation for k-nearest neighbor and neural networks
11/09-11/15	Learning algorithm design for SVM
11/16-11/22	Evaluation and validation, results discussion
11/23-11/29	Write final report
11/30-12/08	Revise final report

Github link: https://github.com/HaotianShi/CS532