

Laboratory Report

Wang Kou¹, Xu Jiasheng¹, Li Junhui¹, Ji Jia¹, Chen Xinrui¹, and Hu Yuxing¹

⁺these authors contributed equally to this work

ABSTRACT

Dozens of algorithms including KNN, SVM, Decision Tree, are applied with MNIST data. The data was download from the MNIST official website. After running the program, There are four different accuracies resulted from different algorithms and different parameters. The following pages describe how we preprocess our data and apply the algorithms.

Introduction

2.1 KNN The proximity algorithm, or K-Nearest Neighbor classification algorithm, is one of the simplest methods in data mining classification technology. The so-called K nearest neighbor is the meaning of k nearest neighbors, saying that each sample can be represented by its nearest k neighbors. The core idea of the kNN algorithm is that if the majority of the k most neighboring samples in a feature space belong to a certain category, the sample also belongs to this category and has the characteristics of the samples on this category. In determining the classification decision, the method determines the category to which the sample to be divided belongs according to only the category of the nearest one or several samples.

2.2 SVM In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

2.3 Decision Tree A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

2.4 Bayes n probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule, also written as Bayes's theorem) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer, compared to the assessment of the probability of cancer made without knowledge of the person's age.

Results

As Showed in Figure 1,2,3

Discussion

For KNN, we mainly work on preprocess data and find out which parameter has the best accuracies. The K-value of 10 and the weights of distance have been proved that they are two best parameters. As the section 3.1 demonstrated, we have the highest accuracy of 95.11% with the total correct number of 5707. But KNN has the longest running time about 15 minutes. For SVM, the running time is shorter comparted to the KNN, which is about 10 minutes. But SVM has relatively low accuracy: 91.27% with 5476 total correct results. For decision tree algorithm, totally 10000 test set data has been tested. This time we have accuracy of 92.67%, which is higher than SVM algorithm. But Decision Tree has a problem that it is unstable. Each time we have different result. Some of them is fairly low and reaches the bottom line about 50% of accuracy. The reason is still puzzled and we are trying to figure it out. We have not solved Bayes algorithm now. According to our experiments, we find that the KNN algorithm is better than Decision Tree, and Decision Tree is better than SVM. But Decision Tree has a problem of instability.

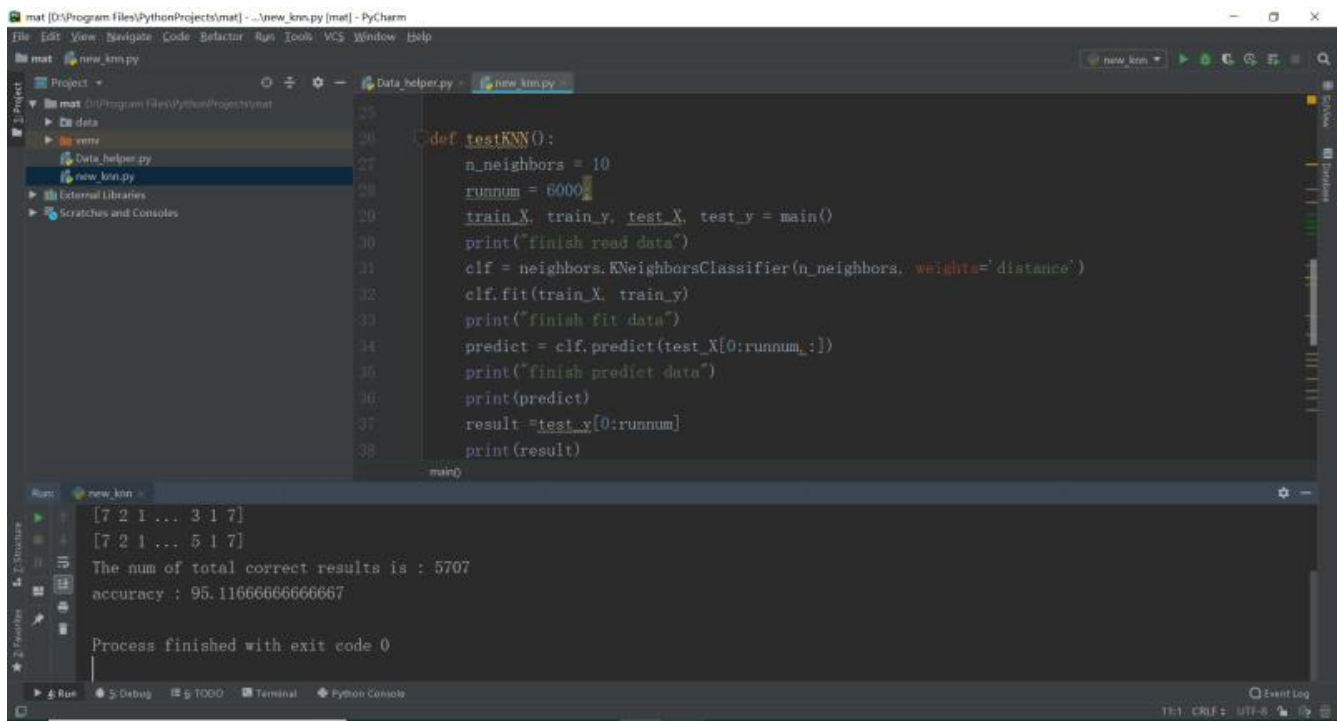


Figure 1. KNN

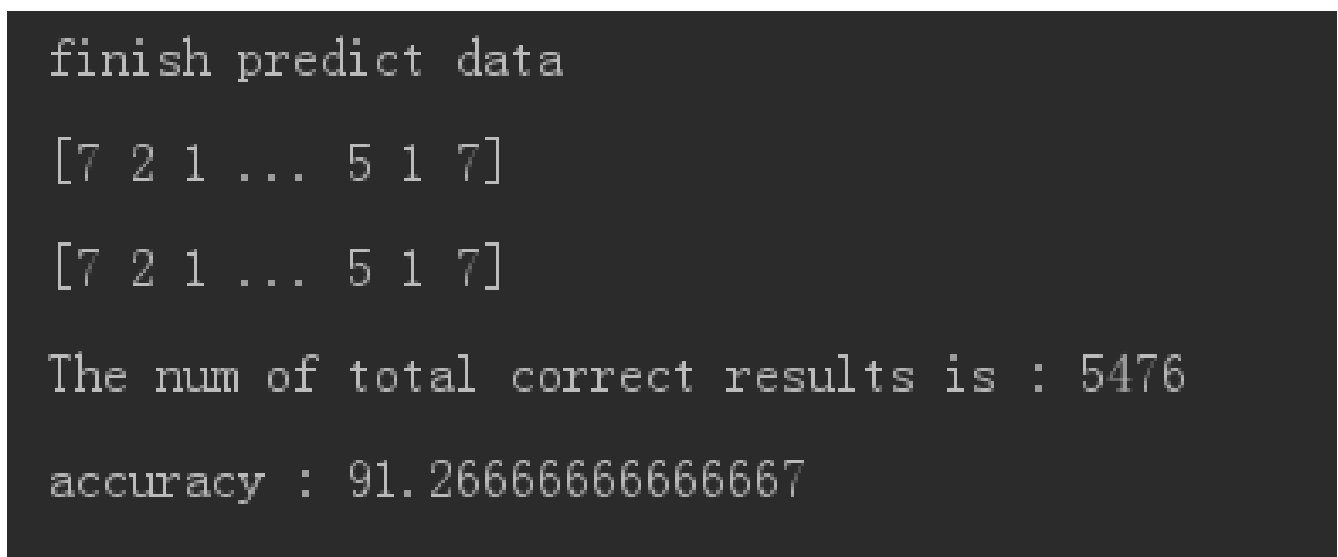
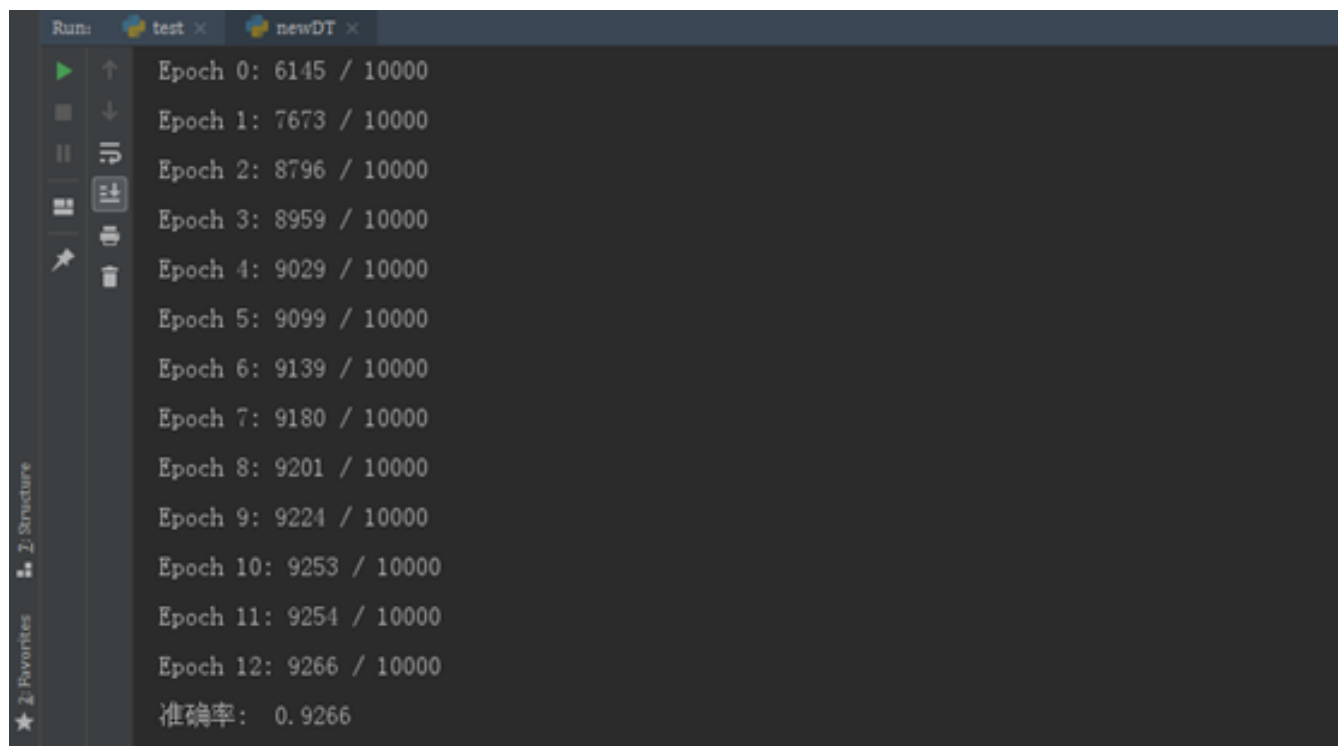


Figure 2. SVM



```
Epoch 0: 6145 / 10000
Epoch 1: 7673 / 10000
Epoch 2: 8796 / 10000
Epoch 3: 8959 / 10000
Epoch 4: 9029 / 10000
Epoch 5: 9099 / 10000
Epoch 6: 9139 / 10000
Epoch 7: 9180 / 10000
Epoch 8: 9201 / 10000
Epoch 9: 9224 / 10000
Epoch 10: 9253 / 10000
Epoch 11: 9254 / 10000
Epoch 12: 9266 / 10000
准确率: 0.9266
```

Figure 3. DCT