Class Handwritten Digits(Modified MNIST)

Full name:

Number:

Kaggle username:

# Introduction

The goal of competition about detection of handwritten digits is to design a machine learning algorithm that can automatically classify images. This dataset has been generated from the popular MNIST dataset. All the images (including held out test samples) are of dimension 56\*28，the number of classes is 19(0~18). The training and held-out test set for this data consists of 50,000 samples and 10,000 samples respectively. The goal is to automatically classify the digits into one of 19 classes from 0 to 18.

A Logistic Regression Model is contrusted from scratch for training data, and on which the accuracy on the test data reached 21.5%.

Then a Random Forest Model is trained, and on which the accuracy on the test data reached 72.0%.

Finally, we construct a simple CNN model, and the accuracy reach 91.9%.

# Feature Design

For the experimental data, each piece of data consists of 56\*28 features with a total of 1568 dimensions. The characteristics of the original training data are shown as follows:

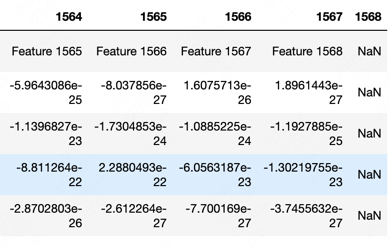
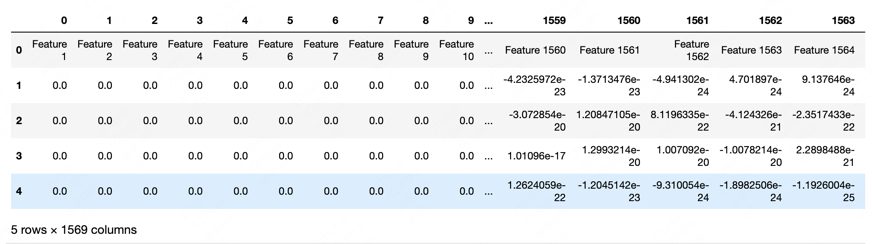


Figure 1.2 Features of training data display

As can be seen from Figure 1.2, the training data has a total of 1569 columns, and the value of the last column is "NaN". When selecting features, the last column should be removed. At the same time, the value of some dimension features is 0.0. For these dimensions, I can choose not to learn the corresponding weights to achieve the effect of reducing the amount of model parameters. Therefore, count the dimensions with dimension value 0.0 in all training data and remove them from the training data. Finally, 1495-dimensional features are obtained for the training data. The most important thing is to keep features of the same dimension as the training data for the testing data.

# Algorithms

## Logistic Regression

Logistic Regression is a commonly used classification model. The essence of Logistic is to estimate model parameters according to the distribution of data.

By analyzing the classification requirements of this competition, the model needs to be able to achieve multi-category classification. For multi-classification problems, the softmax function is used as the activation function, the cross-entropy loss function is used to train the model, and the stochastic gradient descent algorithm is used to optimize the parameters.

The Softmax formula is:

The cross entropy loss function formula is:

And is a one-hot vector，When the current category is the category corresponding to the training data, the value is 1，otherwise is 0.

The parameter update formula is:

## Random Forest

Random forest uses decision trees as weak learners.

The three most important concepts of Decision Tree are Information, Entropy, and Information Gain.

The Information formula is:

Entropy represents the uncertainty of things, the greater the entropy, the greater the uncertainty.

In the Decision Tree algorithm, Information Gain is an index used to select features. The greater the information gain, the better the feature selectivity.

When classifying a sample, each decision tree needs to classify the sample, and vote the classification results of several weak classifiers to form a strong classifier.

## CNN

Considering the correlation and translation invariance of adjacent elements of the image, we design a simple CNN model for experiments. This model consists of two convolutional layers and two fully connected layers.

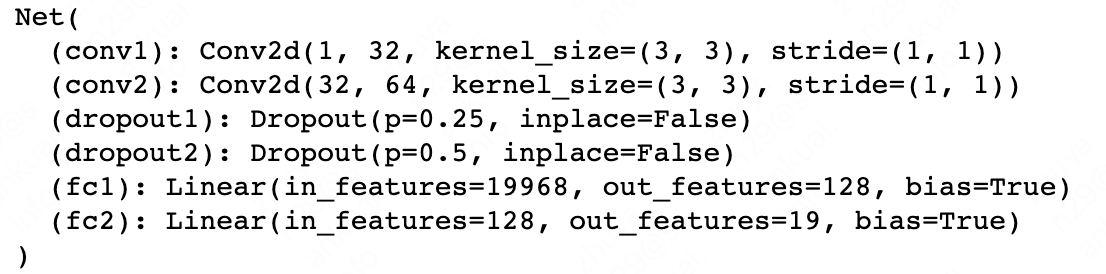


Figure 3 a simple CNN model structure

# Results

## Logistic Regression

Here are some experimental data obtained on this model, presented in tabular form:

Table 1 Experimental data on Logistic Regression

|  |  |  |  |
| --- | --- | --- | --- |
| Epoch | Learning rate | Accuracy on training data | Accuracy on testing data |
| 500 | 0.25 | 0.2477 | 0.213 |
| 2000 | 0.28 | 0.2792 | 0.215 |

The epoch parameter represents how many times the training data has been trained. As the epoch value increases, the model fits better on the training data, and the effect on the test data is also improved.

The learning rate parameter represents the step size of gradient descent, when the value is too small, the update speed is too slow. When the value is too large, it is easy to rush out of the optimal solution range, resulting in a very irregular decrease in the loss value.

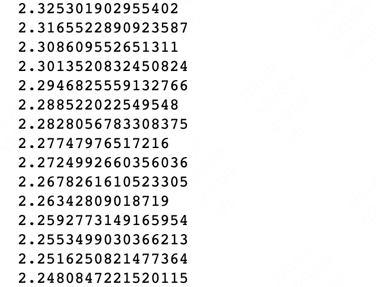


Figure 4 Loss drop when learning rate=0.28

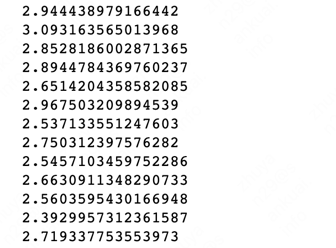


Figure 5 Loss drop when learning rate=0.5

## Random Forest

Here are some experimental data obtained on this model, presented in tabular form:

Table 2 Experimental data on Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
| criterion | N\_estimators | Accuracy on training data | Accuracy on testing data |
| gini | 100 | 1.0 | 0.72 |

## CNN

Table 3 Experimental data on CNN

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Learning rate | Activate | Loss | Optimizer | Epoch | Accuracy |
| 3 | Relu | Cross-entropy | Adadelta | 50 | 0.919 |

# Discussion

The features of handwritten digits are composed of 56\*28-dimensional features, and handwritten numbers are represented by images. Digital images have the following characteristics:

1. There is a certain correlation between adjacent features;
2. Digital images can be recognized whether they are translated, rotated, or scaled .

The Logistic Regression and Random forest used in the experiment do not take into account the above two characteristics, convolutional networks used for experiments can achieve more accurate identification,.

# Statement of Contributions

I hereby state that all the work presented in this report is that of the author.