# presentation

Logistics regression

Logistic regression is a widely used shallow learning model. It mainly solves the problem of binary classifications. The amount of calculation during classification is tiny, the speed is fast, and the storage resources are low. The result of logistic regression is the probability, which helps one see how well a sample fits. When the multi-classification problem occurred in this project, we used SoftMax regression to replace it.

Random Forest

Random forest is very suitable for processing datasets with multiple features. It contains multiple decision tree classifiers. Trees are independent of each other during training, which can significantly improve the training speed. When the data set has a simple structure and fewer features, we use decision tree regression instead of Random Forest regression to enhance efficiency.

MLP

MLP is a neural network composed of fully connected layers with at least one hidden layer, and an activation function transforms the output of each hidden layer. Through Forward propagation and Backpropagation, iterative weight parameters have achieved the purpose of training the model.

CNN

CNN is a type of feedforward neural network that includes convolution calculation and has a deep structure. The deep feature value is extracted through the operation of convolution. It has the characteristics of a shared convolution kernel, which allows it to process higher-dimensional data, such as pictures. Like MLP, it is prone to overfitting.

german Traffic sign p1:

As we all know, one of the most popular application areas of artificial intelligence is **self-driving**. Traffic sign recognition is one of the problems what we need to be solved urgently.

Its a picture dataset of **43 street signs** in Germany. Each picture is a 32x32 three-channel color image. We use OpenCV to **convert** them to gray images to reduce the dimension. CNN and MLP models use three-channel color images, RNN uses gray one, and other shallow learning methods have to reshape the image to flatten dimensions 1024. The picture on the right is what looks like in jupyter notbook, but it is a black and white image actually.

\*\*\*\*\*

because(Since) the dataset is belong to color image, using **shallow** learning methods need to **reduce** the dimension (of the image). It will make the image lose a lot of information. Its the main reason why the **poor results** in random forest and SoftMax regression.

Using the models of deep learning have been improved obviously. Among them, CNN could **extract region information** of image through convolution to make the highest accuracy. RNN is not suitable for processing image information, so the result is lower than CNN and MLP. On the right is the structure of the CNN model, which uses multi-layers convolution and two max pooling layers. and here we can see form the accuracy map, after the second epoch the tends to **stabilize** .

\*\*\*\*\*

This table is the results of combination models . The accuracy of random forest result **as an extra dimension** is higher than the previous MLP model. It indicates that the probability values come from the shallow learning model **play an active role** in the final classification of deep learning.

We build several models of shallow learning methods **applied to the extracted layer** from deep learning. Here are random forest and SoftMax regression after CNN fully connected layer, the others are random forest and SoftMax regression after CNN second max pooling layer. The accuracy results are much higher than that using shallow learning alone. After the convolution operation, the deep features have **a higher amount of information** and it **helps** shallow learning method to classify. however, the side effect is looks have little a bit overfitting.

\*\*\*\*\*

Predicting a Pulsar Star：

Next we are talking about Pulsar star dataset. The data records 8 feature values, including mean, standard, and so on. They are using these features to predict whether a star is a pulsar star or not. It is a **binary** classification issue typically. there are1639 pulsar star in the dataset.

As can be seen from this table, some value range of features are quite large, while some value ranges are small. in this case, Normalized the dataset could **improve** the **convergence effect** of gradient descent. and **here** we use StanderScaler from sklearn .

\*\*\*\*\*

we could see from the table. Due to the simple data structure and fewer features, almost each model can produce satisfactory result. Logistic regression is the fastest model in this case, so In my opinion logistic regression is more suitable for this dataset.

The left picture shows the structure of the MLP model which only use 3 fully connected layers. And look the model accuracy map, MLP model can reach more than 97% after the first epoch. It is **high-efficiency** even using deep learning method.

We could get a conclusion that both deep learning and shallow learning models could get good results for a small number of features dataset.

\*\*\*\*\*

here is the combination models. In the model of logistic regression as **extra features**, the accuracy is **slightly lower** than the previous MLP model. It might be the probability of Logistic Regression results is **not correct enough**. In other words, The result of shallow learning may **improve** the accuracy of deep learning. On the other hand, it may **add noise** to deep learning models at the same time.

when using Logistic Regression and random forest on last MLP layer, the results are almost the same with previous models. The reason is that the structure of the dataset is simple . The **previous** research accuracy is quite high already, and there no more space to improve in those models.