Machine learning classification for trading signals

content

Introduction

In market trading, we want to make profits as well as we can, so we have to predict the up-or-down trends in the future accurately. As we all know, there are many signals which are helpful for people to predict the future of market. If we are lucky to find such important signals and discover relations between its and trend of the market before others, we can enrich ourselves.

Now not only investors in the marker but researchers are do their best searching the optimized solutions including the most valuable trading signals and the optimized trading strategies. In this report, I try several statistical models and machine learning models including logistic regression, svm, xgboost and artificial neural networks. By building different models and comparing the performances in validation data, we final a relatively model as our final choice. But the way to find the optimized solutions is still so long, we need to put more energy and time in it.

Data

The dataset of this report contains two parts, namely the 'train' part, and 'test' part. According to the variable named 'data_type' in 'test' part, we cut the original 'test' part and get the 'evaluation' part which data_type is 'validation' and the 'test' part which data_type is not 'validation'. Just as its name, the 'train' part is for training the model, 'evaluation' part is for validation and 'test' part is for testing the performance of model. All three parts have 50 features which has been encoded and standardized for confidential reasons thus we have no idea its real meaning. Also, all are continuous variables and between 0 and 1. One of our tasks is to find the important variables as the most valuable signals which can tell us how the market will change in the future. Compared to 'test' part, 'train' part and 'evaluation' part have target variable which is binary variable showing the up-and-down of the market.

For this classification task, we need to check the distribution of target. Correspond proportion of target are showed as follow, the positive samples and negative samples are balanced thus we can use it directly without additional data preprocessing.

data	Proportion of target 1(%)
Train data	50.01908426082841
Validation data	49.9928678410955

Methods

In this situation, the target we pay attention to is a binary variable, thus it is a classification problem. Basically all the classification model can be applied in this situation if we don't mind its performance in the dataset. As we have not done such data mining before, we try several methods of machine learning which are common and we are familiar with.

Firstly we try logistic regression. Compared to linear regression, the response variable is binary. Here we regard all 50 feature variables as explanatory variables to build logistic regression model. Compared to other model we would mention below, logistic regression are good at explaining the relation between corresponding features and target variable. We can find out which features are positive with target variable and which are negative clearly. Also, it is obvious to drop out such features that have no impact on target variable. Then we try a model named svm with gaussian kernel. The reason we try svm is that in small data set, svm always perform better than other classification model. It is very robust so outliers have little impact on it during training. Next, we build a xgboost model by setting its max depth as 5 and number of boosting iterations as 3 and so on. Xgboost is a method of boosting. It would train many classifiers and merge all the classifiers. Finally, we try a method of artificial neural networks by build two dense layers, the first with 64 units and 'relu' activation, the next with 2 units and 'softmax' activation. Also , we set the batch size

as 20 and epoch as 20 and apply the ‘adam’ training algorithm. In order to obtain a best model during training, we use the early_stop mechanism, it would stop training when the loss of validation data don’t become lower in 3 epochs. Artificial neural networks are popular not only in industrial community bur also in academic community. it represents deep learning and always have best performance in many situations, like image processing, natural language processing and so on. It have a great power on digging valuable features from data. But it works like a black box so it lacks the ability to explain the essence.

All the model mentioned above are trained on ‘train’ part data. After training, we evaluate its performance including log loss and accuracy on ‘evaluation’ part data. All the performance on validation data are showed as followed. From the table below, it is easy to see that all the index between these models are close to each other. In detail, svm model behaves worst on the validation data, its corresponding log loss is largest and its corresponding accuracy is min.

The other two model logistic regression and xgboost are close to each other, logistic regression is better on log loss but xgboost has a higher accuracy. It is better than svm but lower than ann. What we need to pay attention to are artificial neural networks and LightGBM, one has the smallest log loss and one has the highest accuracy.

model	information	Log loss	accuracy	comment
Logistic regression		0.696424225193728	0.5004814207260537	medium
SVM	Gaussian kernel	0.7287187402848114	0.49511447115041723	Worst performance
xgboost		0.7005587471630452	0.5044575993153128	medium
ann	Dense network	0.693500896144352	0.5107517295485343	Maximum accuracy
LightGBM		0.6931393482957735	0.5044575993153128	Minimum log loss

Finally, we use trained models to predict the targets on the 'test' part data and obtain corresponding probability of target is 1. All the predictions are saved in a excel. Its columns represent different models and its rows represent correspond probability of target is 1 on the 'test' part data.

Discussion

Before our experiment, we conjecture that the method of deep learning would perform better than other traditional method of machine learning. The result is as our conjecture. The artificial neural networks can extract important features from 50 original features in a unusual way which can't be done by human. Thus its has strong ability to discover the essence of nature than other method of machine learning can get.

All our models here can be improved by adjusting the inner structure further. All

the models are built and trained in a usual way with default parameter. So, we can improve the performance of models by adjusting corresponding parameters. For example, we can increase the number of units and the number of layers in artificial neural networks. Also, we can try other training algorithm and batch size and so on. For machine learning model svm, we can use other kernel or change the parameter of penalty.

In conclusion, the model of artificial neural networks works best in this dataset but not reach our requirements. It is a long way to find the optimized solution, not only the most valuable signals but optimized trading strategies.