**武汉大学计算机学院**

**本科生课程设计报告**

**灾难中乘客生还概率预测**

专 业 名 称 ：软件工程

课 程 名 称 ：商务智能

指 导 教 师 ：朱卫平

班级 ：卓越二班

学 生 学 号 ：2016302580320

学 生 姓 名 ：任思远

二○壹陆年12月

**郑 重 声 明**

本人呈交的设计报告，是在指导老师的指导下，独立进行实验工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本设计报告不包含他人享有著作权的内容。对本设计报告做出贡献的其他个人和集体，均已在文中以明确的方式标明。本设计报告的知识产权归属于培养单位。

本人签名： 日期：

Abstract

This project talks about the prediction on potential survivors in disasters, so that rescue plan can be made according to the result to reduce casualties to the utmost extent.

At first, feature engineering and data cleaning is conducted to prepare for the input for neural network. The feature selection matters a lot and will impact on the model significantly. Then construct the neural network with a simple model (two-level layers using Relu as active function and Softmax cross entropy loss and Adam Optimizer). Finally train and test on AI Studio using PaddlePaddle.

Keyword: AI Studio; PaddlePaddle; survival prediction; Neural Network; Feature engineering;

Table of Contents

Background ------------------------------------------------------------ 5

Demand Analysis ----------------------------------------------------- 5

Solution ---------------------------------------------------------------- 6

Core implementation ------------------------------------------------- 6

1. Feature engineering and data cleaning ---------------------------- 6
2. Neural Network Constructing -------------------------------------- 10
3. Training ---------------------------------------------------------------- 11
4. Test --------------------------------------------------------------------- 12

Data Sources --------------------------------------------------------------- 12

Run on AI Studio --------------------------------------------------------- 13

Parameter Specification ---------------------------------------------- 16

Discussion and Experience ------------------------------------------ 17

References ------------------------------------------------------------------ 17

Background

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. This sensational tragedy shocked the international community and led to better safety regulations for ships. For this kind of disasters, it’s critical to perform in-time rescue with high success rate. However, it’s often to be proved that the limited rescue resources cannot cover the whole disaster site since the outside has no information about the situation inside.

In such case, the potential survivor information will be helpful for us to conduct the rescue plan. For example, if we know the passengers in cabin ‘A’ have more survival possibility, we can schedule more rescue power and resources aiming to that cabin, so as to reduce casualties to the utmost extent.

Demand Analysis

In this project, the task is to predict whether the passengers can be alive in the disaster. I choose the Kaggle competition [Titanic: Machine Learning from Disaster](https://www.kaggle.com/c/titanic), in which the task is to analyze which kind of people are likely to survive. The detailed information about each person include ‘Ticket class’, ‘Sex’, ‘Age’, ‘number of siblings / spouses aboard the Titanic’, ‘number of parents / children aboard the Titanic’, ‘Ticket number’, ‘Passenger fare’, ‘Cabin number’, ‘Port of Embarkation’.

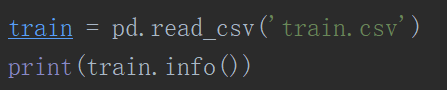
Solution

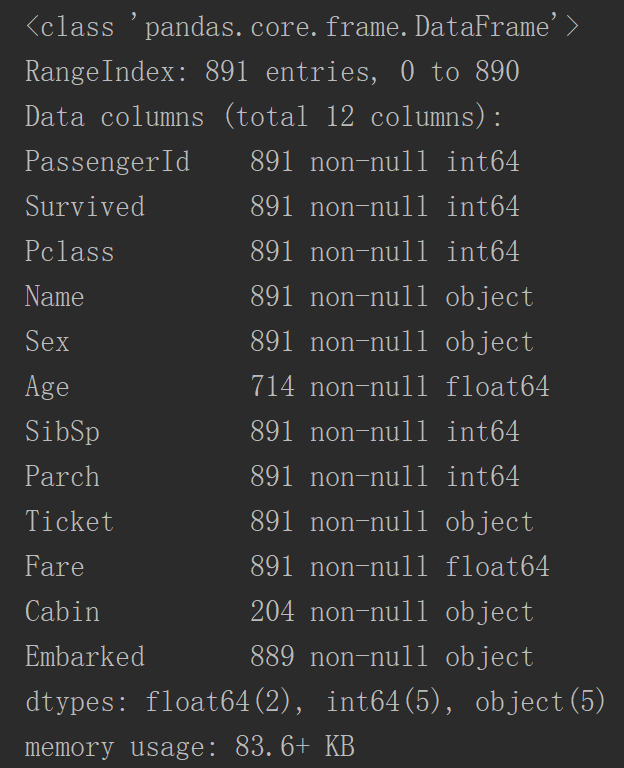
The main idea is to construct a two-level layers neural network, **RELU** being active function, **cross verification** and **Adam Optimizer**.

Core Implementation

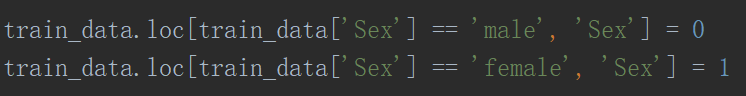
1. **Feature engineering** and **Data Cleaning**

Take a brief look on the training data, what we have learned is that attribute ‘Age’ is partly absent, attribute ‘Cabin’ is mostly absent so that it’s might be of no use, ‘Embarked’ has only 2 data absent.





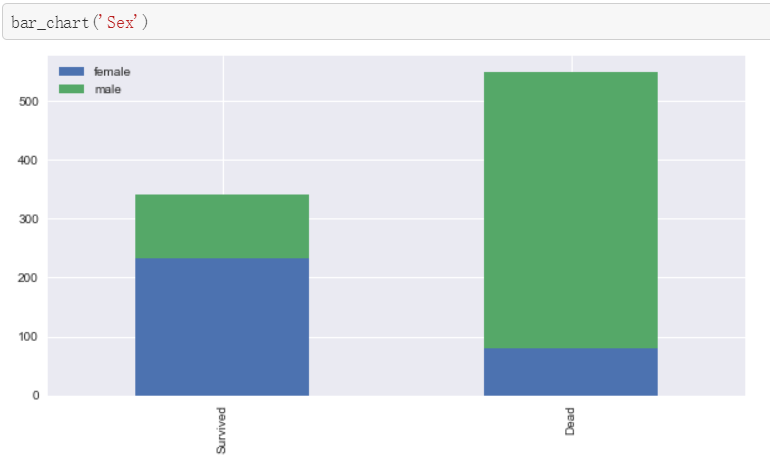
As for ‘Sex’ attribute, we encode ‘male’ as 0, and ‘female’ as 1.



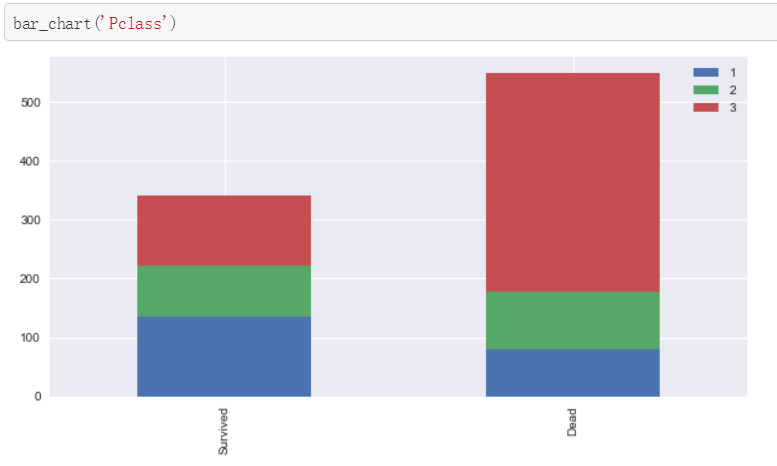
As for ‘Embarked’ attribute, there are 3 irrelevant values, so it should be imposed by one-hot encoding.

As for ‘Age’ attribute, we use **RandomForest** to predict the absent values.

Feature Selection:

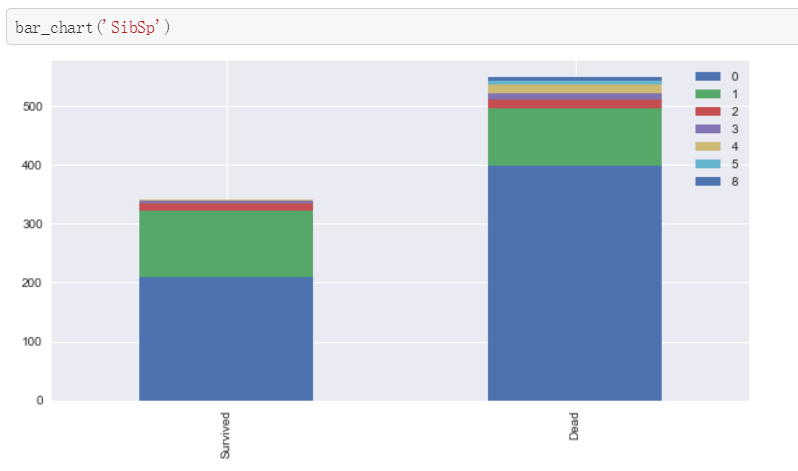


The Chart confirms **Women** more likely survived than **Men**.



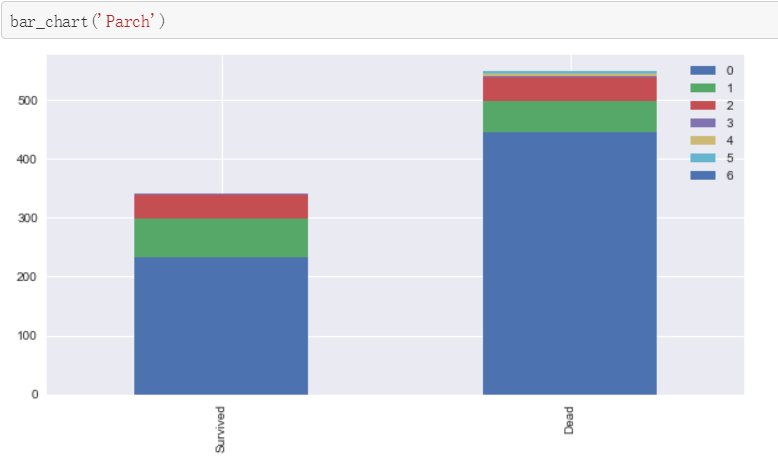
The Chart confirms **1st class** more likely survived than other **classes**.

The Chart confirms **3rd class** more likely dead than other **classes**.



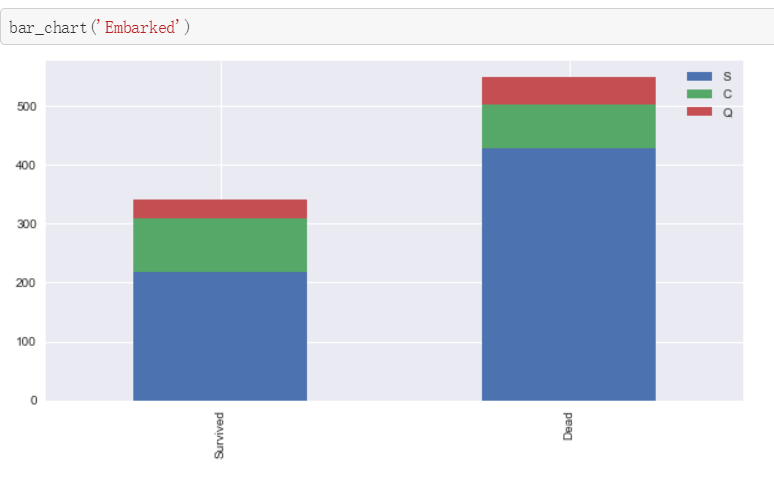
The Chart confirms **a person aboarded with more than 2 siblings or spouse** more likely survived.

The Chart confirms **a person aboarded without siblings or spouse** more likely dead.



The Chart confirms **a person aboarded with more than 2 parents or children** more likely survived.

The Chart confirms **a person aboarded alone** more likely dead.

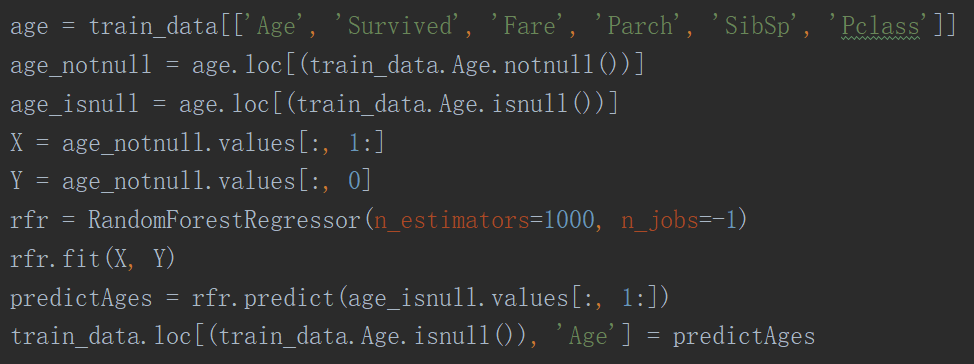


The Chart confirms **a person aboarded from C** slightly more likely survived.

The Chart confirms **a person aboarded from Q** more likely dead.

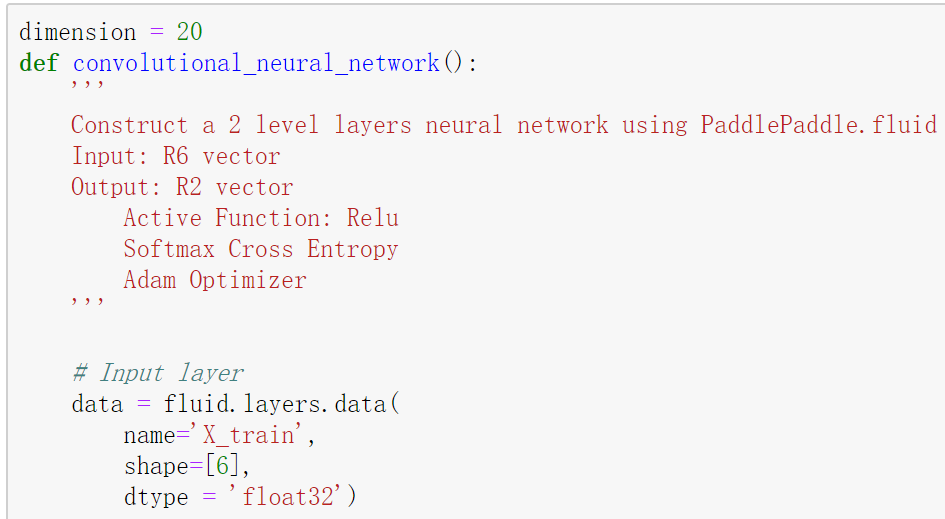
The Chart confirms **a person aboarded from S** more likely dead.

From above information, we pick 6 dimensions about each person to serve as input data to predict the survival rate.



1. **Neural Network Constructing**

The target is a two-level layers neural network.



Structure for neural network:

Construct a 2 level layers neural network using **PaddlePaddle.fluid**

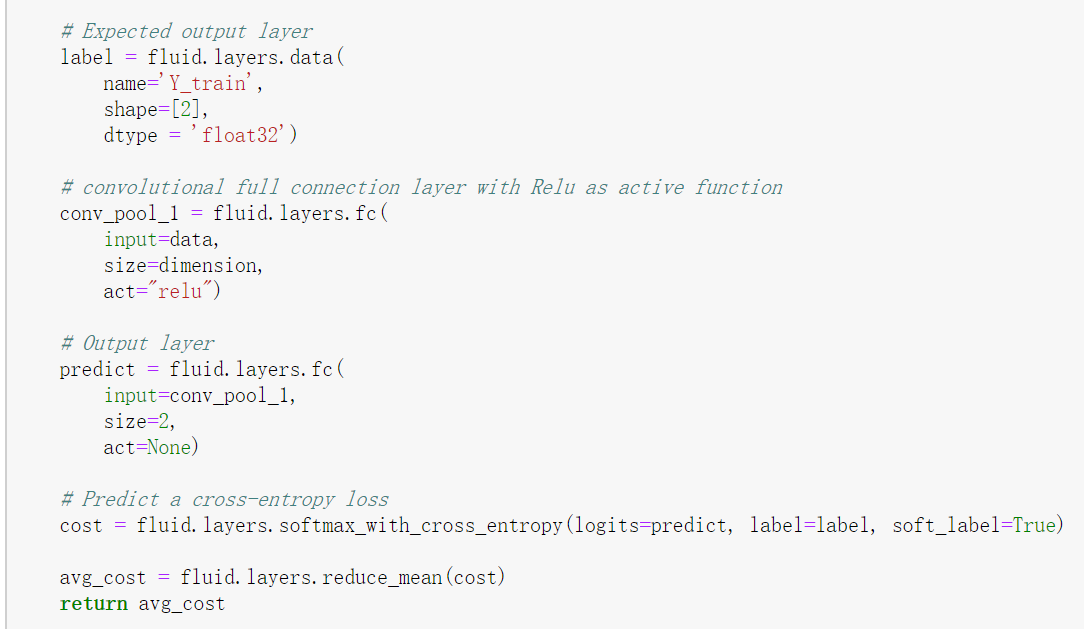
Input: R6 vector

Output: R2 vector

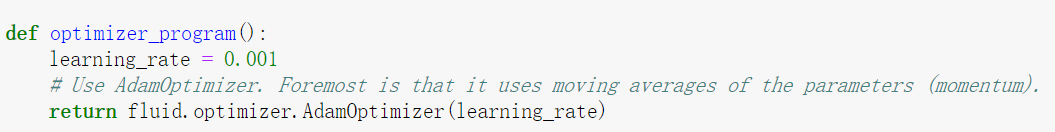
Active Function: Relu

Softmax Cross Entropy

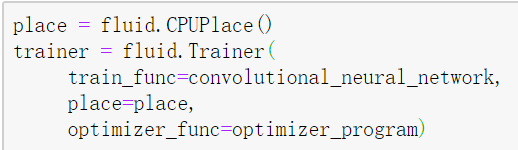
Adam Optimizer

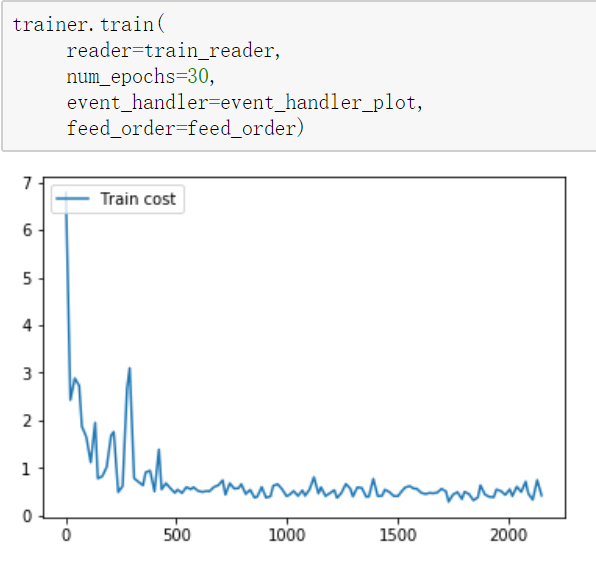


Use AdamOptimizer. Foremost is that it uses moving averages of the parameters (momentum).



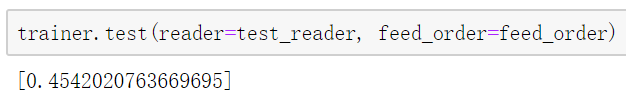
1. **Training**





The result is acceptable, then make a testing on the model.

1. **Test**



The cross entropy loss converges to 0.

Data Sources

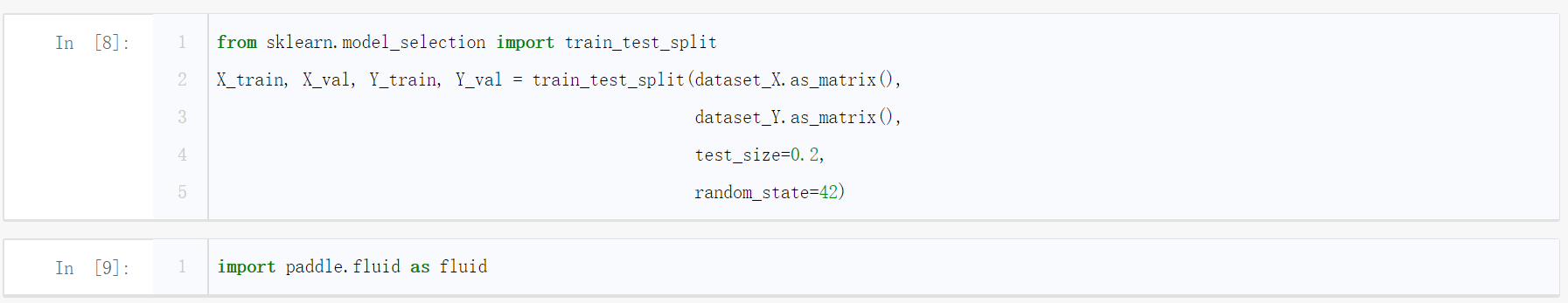
Data is from: [Titanic: Machine Learning from Disaster - Data](https://www.kaggle.com/c/titanic/data), including 2 files: “train.csv”, “test.csv”.

Run on AI Studio

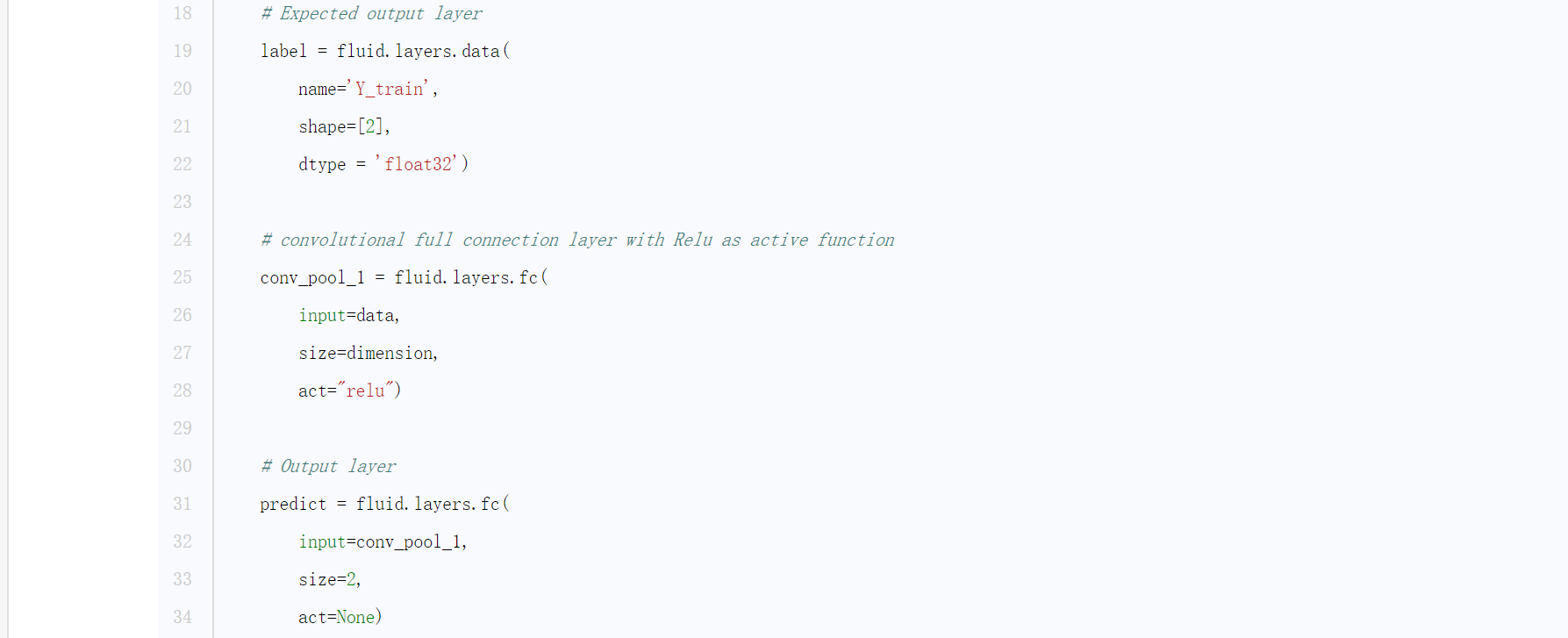
The project is published at: <http://aistudio.baidu.com/#/projectdetail/33985>

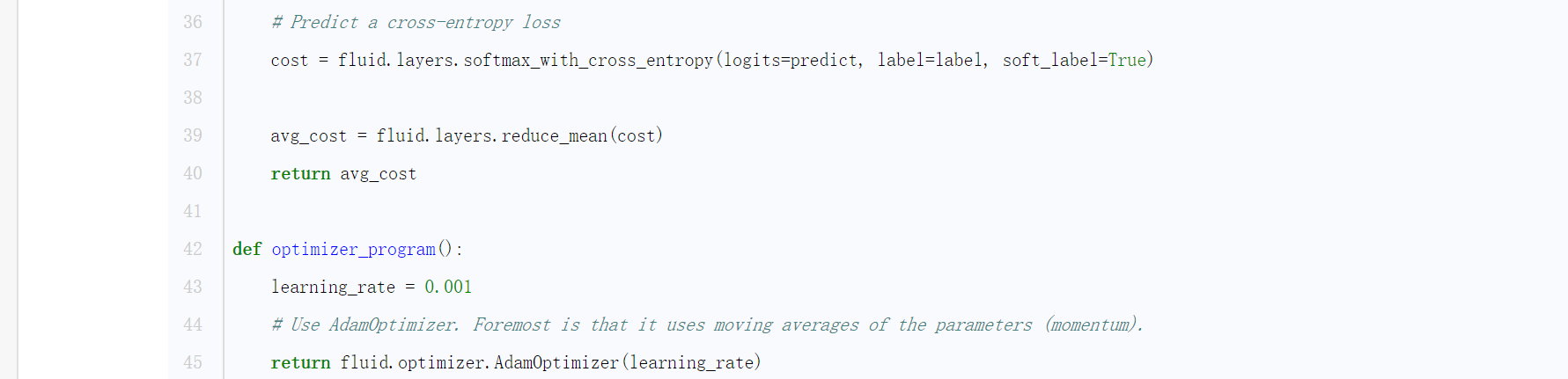




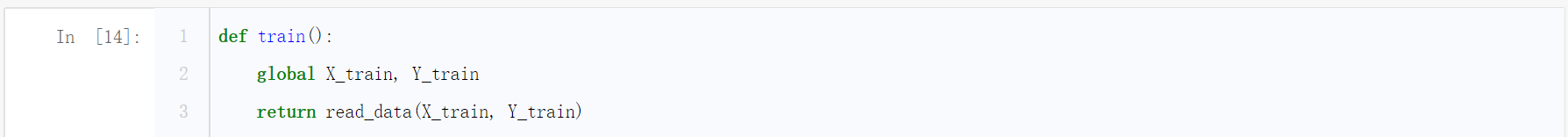




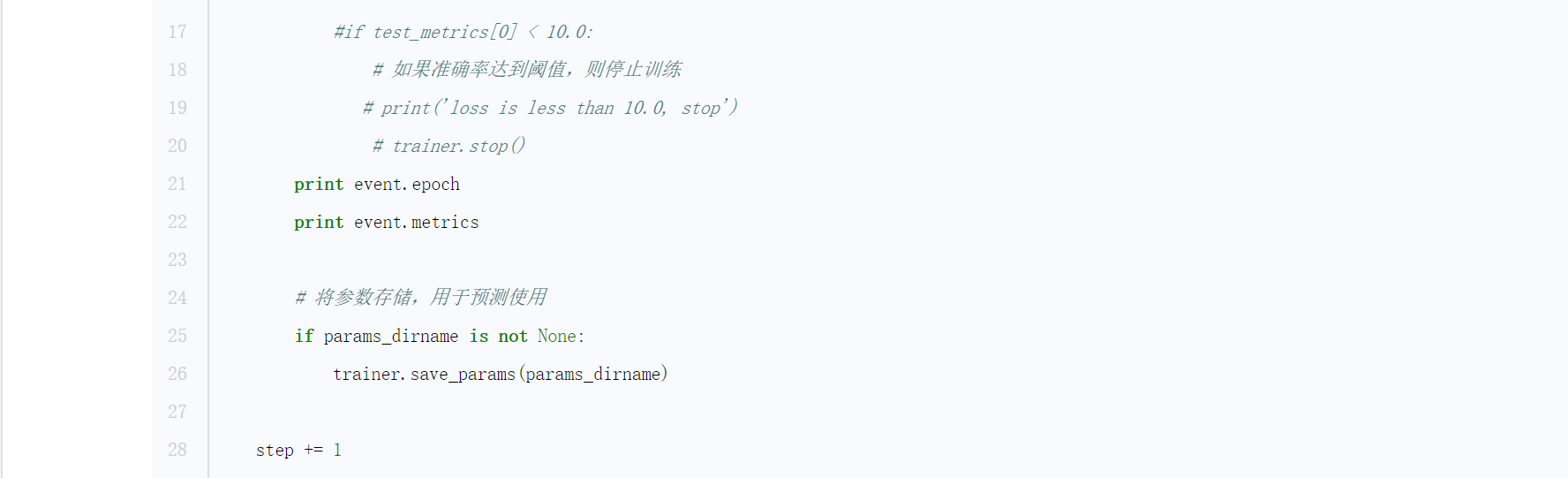


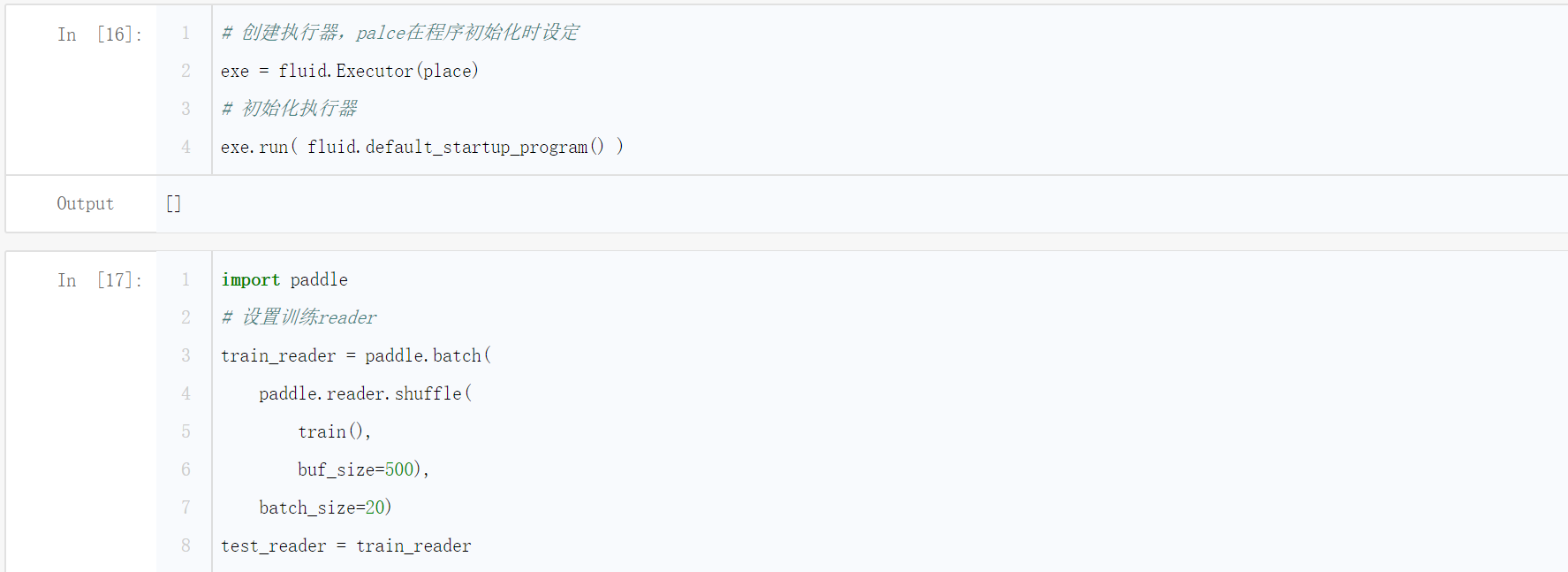


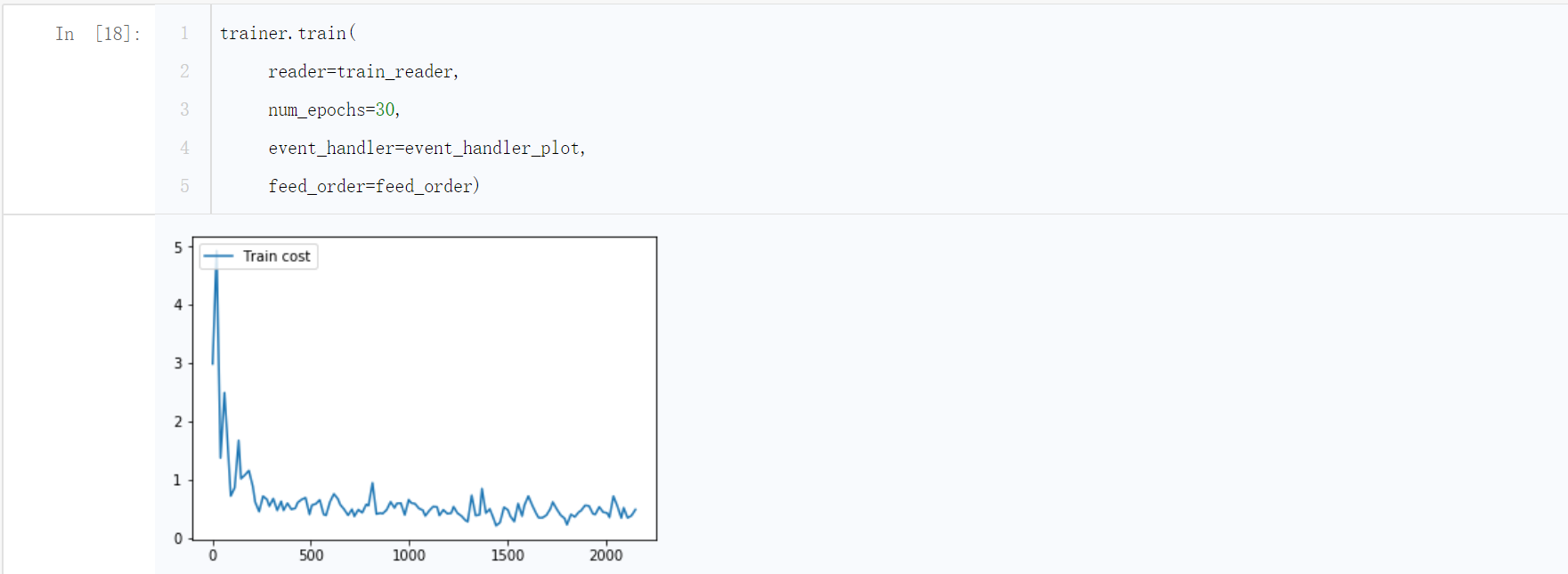


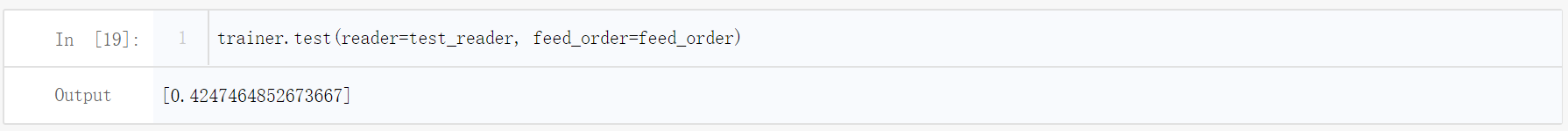




















Parameter Specification

Neural Network structure: two-level, 6 -> dimension -> 2

Active function: Relu

Loss function: Softmax and Cross Entropy

Optimizing function: Adam Optimizer

Data input: shuffled by **paddle.reader.shuffle** and make a batch by **paddle.batch.**

Train Epochs: 30

Discussion and Experience

1. Feature engineering and data cleaning is so important even more than model itself sometimes.
2. The epistemology of data matters.
3. Model ensemble performs better.
4. Parameters in convolution network worth spending a period of time to figure out.

References

[1] Difference between GradientDescentOptimizer and AdamOptimizer (TensorFlow)?<https://stats.stackexchange.com/questions/184448/difference-between-gradientdescentoptimizer-and-adamoptimizer-tensorflow>

[2] PaddlePaddle 使用文档 http://www.paddlepaddle.org/documentation/docs/zh/0.14.0/new\_docs/beginners\_guide/index.html